Proceedings

THIRTY-EIGHTH ANNUAL ALLERTON CONFERENCE ON COMMUNICATION, CONTROL AND COMPUTING

October 4 - 6, 2000

Allerton House, Monticello, Illinois
Sponsored by the
Coordinated Science Laboratory and the
Department of Electrical and Computer Engineering of the
University of Illinois at Urbana-Champaign

PROCEEDINGS

THIRTY-EIGHTH ANNUAL ALLERTON CONFERENCE
ON COMMUNICATION, CONTROL, AND COMPUTING

VOLUME II

3/13/2001

NOTICE: **Return or renew all Library Materials! The *Minimum* Fee for each Lost Book is $50.00.**

The person charging this material is responsible for its return to the library from which it was withdrawn on or before the **Latest Date** stamped below.

Theft, mutilation, and underlining of books are reasons for disciplinary action and may result in dismissal from the University.
To renew call Telephone Center, 333-8400

UNIVERSITY OF ILLINOIS LIBRARY AT URBANA-CHAMPAIGN

FEB 14 2004
FEB 14 A.M.
FEB 14 2004
SEP 23 2005

L161—O-1096

The

THIR
ON C

R.S. Sreenivas
Douglas L. Jones
Conference Co-Chairs

Conference held
October 4 - 6, 2000
Allerton House
Monticello, Illinois

Sponsored by
The Coordinated Science Laboratory
and
The Department of Electrical and Computer Engineering
of the
UNIVERSITY OF ILLINOIS
at
Urbana-Champaign

Equalization of Space-Time Coded Multi-antenna Systems

Harish Viswanathan
Lucent Technologies Bell Laboratories
Holmdel
NJ 07733
harishv@lucent.com

Abstract

Space-time coding techniques have been introduced to achieve maximum transmit diversity and coding gains and/or to increase throughput when multiple antennas are available at the transmitter and receiver. We consider the application of these space-time codes to multiple-input multiple-output (MIMO) frequency-selective fading channels wherein an equalizer is required to suppress the inter-symbol interference. We present novel linear and decision feedback equalization schemes that can be used in combination with the space-time decoders designed for flat-fading channels to realize the diversity gains. These equalization schemes differ from traditional equalization schemes in that they are designed to suppress only the inter-symbol interference, but not the mutual interference between the different transmit antenna signals which is handled by the maximum likelihood space-time decoder. When multiple receive antennas are available the linear equalizer derived herein generalizes the notion of principal ratio combining to frequency-selective fading allowing for tradeoff between performance and complexity.

1 Introduction

The capacity of wireless communication systems over fading channels is significantly enhanced by the use of multiple antennas at the transmitter and receiver [5]. New space-time codes have been proposed for use in systems with multiple transmit antennas to achieve diversity gains [1], [10] as well as for increasing throughput [4], [8]. Most of these schemes have been proposed and studied under flat-fading channel conditions. In channels with inter-symbol interference (ISI) equalization is necessary for successful demodulation of the transmitted information bits. Equalization for traditional single-input single-output (SISO) systems and also single-input multiple-output systems has been the subject of study for several years. These equalization schemes are concerned primarily with the successful demodulation of a single data stream from the received signal corrupted by inter-symbol interference and additive noise. More recently there has been some research on equalization techniques for multiple input multiple output (MIMO) systems [2], [7]. Nevertheless, the focus is often only to recover a single desired data stream of interest from the received signal which is corrupted by self-interference in the form of ISI and mutual interference from the other inputs. The techniques are targeted towards interference cancellation and equalization of the self-ISI. In order to achieve the

2000:2

full diversity guaranteed by the space-time codes it is necessary to jointly demodulate and decode the multiple inputs. In this paper, we present schemes for equalization of space-time coded systems and some novel techniques for equalization and demodulation of multiple streams in a Bell Labs Layered Space-Time (BLAST) [6] or MIMO system.

We consider reduced complexity techniques such as linear and decision feedback equalization. Maximum likelihood sequence estimation (MLSE) is considered in [11]. We formulate appropriate optimality criteria to derive these equalizers that are different from traditional optimality criteria such as minimum mean squared error (MMSE). The optimal solutions turn out to be the principal eigen vectors of the signal and interference matrices and reduce to the traditional equalizers for SISO systems. The paper is organized as follows. In Section 2 we present a brief description of space-time coding techniques. In Section 3 we present the received signal model for the MIMO frequency-selective fading channel. In Section 5 we present reduced complexity algorithms for equalization of space-time coded systems. In Section 6 we present some simulation results to illustrate the performance of the equalizers for space-time coded systems.

2 Background on Space-time coding

Space-time coding is a channel coding and modulation scheme for multiple antennas that tries to achieve maximum possible transmit diversity and coding gain. Alternately, when multiple antennas are available at the terminal as well, instead of coding for maximum diversity one can consider coding schemes that enhance the throughput [6], [8]. These schemes were originally proposed for frequency-flat-fading channels and hence did not require an equalizer. We describe both the trellis based space-time coding for maximum diversity and the BLAST scheme briefly in order to motivate the need for new equalization structures proposed in the next section.

In trellis space-time coding information symbols are mapped to coded modulation symbols based on a trellis that defines the input-output relationship and the ending state for each starting state. In the trellis code for achieving 2 bits/symbol using QPSK modulation with two antennas in [10], at each time instance depending on the state of the encoder and the two input bits a branch having a label $b_1 b_2$ is selected and $\exp(b_1 \pi j/2)$ and $\exp(b_2 \pi j/2)$ are sent simultaneously from transmit antennas one and two. Assuming that the transmitted symbol at symbol period n from the m^{th} antenna is a_n^m, the received signal at antenna p under the flat-fading channel model, is

$$r_p(n) = \sum_{m=1}^{M} h_p^m a_n^m + w^p(n)$$

where the fading coefficients h_p^m are modeled as independent samples of complex Gaussian random variable with mean zero and variance 0.5 per dimension and M, P are the number of transmit and receive antennas, respectively. Maximum likelihood decoding assuming ideal channel state information at the receiver chooses the codeword sequence $a_1^1 a_1^2 \cdots a_1^M \cdots a_l^1 a_l^2 \cdots a_l^M$ which minimizes the metric

$$\sum_{n=1}^{l} \sum_{p=1}^{P} \left| r_p(n) - \sum_{m=1}^{M} h_p^m a_n^m \right|^2 \tag{1}$$

among all possible codewords. The metric calculation can be simplified at the expense of some performance penalty for the case when the number of receive antennas P is large by using the technique called principal ratio combining (PRC) [9]. PRC can be viewed as a generalization of maximal ratio combining at the receiver when transmit diversity is employed. We will later show that the concept of PRC extends naturally to the case when the channel is frequency-selective.

When the number of receive antennas is larger then the number of transmit antennas it is possible to separate the signals from different transmit antennas. Hence one can potentially send independent information from the different transmit antennas and increase the throughput of the link. Such techniques have been proposed in [4], [8]. When the information signals are not encoded as in [6], simple linear processing techniques can be used to recover the information signals. Such techniques have been proposed even for channels with ISI in [2]. We present receiver structures that provide improved performance compared to these schemes.

3 Sufficient Statistics

Consider a system with M transmit antennas and P receive antennas. Let $\{a_n^m\}_{n=-\infty}^{\infty}$ be the sequence of information symbols that are transmitted from antenna m. If $p(t)$ is the transmit pulse shaping filter then the transmitted base-band signal from each antenna is given by

$$s_m(t) = \sum_n a_n^m p(t - nT)$$

where T is the transmission symbol period. Consider propagation through a fading multipath channel with L paths. We will restrict attention to quasi-static channels in which the channel remains static during the duration of a burst. Let $\alpha_{l,m,p}$ and $\tau_{l,m,p}$ be the path gains and delays respectively, of the l^{th} path from transmit antenna m to receive antenna p. Then the base-band received signal at antenna p is given by

$$\begin{aligned} r_p(t) &= \sum_{m=1}^{M}\sum_{l=1}^{L} \alpha_{l,m,p} s_m(t - \tau_{l,m,p}) + w_p(t) \\ &= \sum_n \sum_{m=1}^{M} a_n^m x_{m,p}(t - nT) + w_p(t) \end{aligned} \tag{2}$$

where $w_p(t)$ is the complex additive white Gaussian noise process at receive antenna p, and we have defined

$$x_{m,p}(t) \triangleq \sum_{l=1}^{L} \alpha_{l,m,p} p(t - \tau_{l,m,p}). \tag{3}$$

We assume throughout the paper that the channel gains $\alpha_{l,m,p}$ and the delays $\tau_{l,m,p}$ are known at the receiver. Typically, these are estimated based on a training sequence that is transmitted during each burst. Given that the receiver knows the pulse shaping filter $p(t)$ we can assume that the receiver has knowledge of the equivalent channel filters $x_{m,p}(t)$.

It can be shown that [11]

$$x_m(n) = \sum_{p=1}^{P} \int r_p(t) x_{m,p}^*(t - nT) dt \tag{4}$$

is a set of sufficient statistics for the maximum likelihood decoding of the information symbols. Substituting equation (2) in (4) we can derive the following equivalent channel model for the system under consideration:

$$\begin{aligned} x_m(n) &= \sum_{p=1}^{P} \int \left(\sum_{n'} \sum_{m'=1}^{M} a_{n'}^{m'} x_{m',p}(t-n'T) + w_p(t) \right) x^*_{m,p}(t-nT)dt \\ &= \sum_{n'} \sum_{m'=1}^{M} a_{n'}^{m'} h_{m',m}(n-n') + w_m(n) \end{aligned} \tag{5}$$

where we have set

$$h_{m',m}(n) = \sum_{p=1}^{P} \int x_{m',p}(t) x^*_{m,p}(t-nT)dt \tag{6}$$

$$w_m(n) = \sum_{p=1}^{P} \int w_p(t) x^*_{m,p}(t-nT) \tag{7}$$

Equation (5) represents an equivalent discrete-time channel model that we will use as the starting point to derive the various equalization schemes. The noise covariance matrix is given by

$$\begin{aligned} \{\Sigma(n)\}_{m',m} &\triangleq E\left[w_{m'}(k) w^*_m(k-n)\right] \\ &= \sum_{p=1}^{P} N_0 \int x_{m,p}(t-(k-n)T) x^*_{m',p}(t-kT)dt \\ &= N_0 h^*_{m',m}(n) \end{aligned} \tag{8}$$

4 Equalization for Space-time Decoding

While MLSE is optimal it is often prohibitively complex even for single-input single-output systems especially for channels with large delay spread. For multiple-input systems the complexity grows exponentially with the number of inputs M in addition to the exponential complexity growth with respect to the delay spread of the channel. Hence it is necessary to derive suboptimal algorithms of reasonable complexity without significant degradation in performance. One approach to reduce complexity is to treat all but one of the inputs as interference and use interference cancellation techniques that reject interference from other inputs while also suppressing the inter-symbol interference from the desired input [2], [7]. In addition to suppressing the interference from the other inputs, the front-end equalization filters are also designed to suppress the inter-symbol interference (ISI) in the desired input. Minimum mean squared error (MMSE) linear and decision feedback equalization techniques have been proposed for the front-end filtering. While this technique is suitable for the BLAST/MIMO type schemes where independent streams are transmitted from the different transmit antennas it does not appear to be suitable for space-time coded systems where the goal is to achieve transmit diversity by simultaneous transmission from the multiple transmit antennas. This is especially true for the case when the number of receive antennas is less then the number of transmit antennas in which case it is not possible to separate the multiple input signals through linear processing. We present equalization schemes where the equalizer cancels only the ISI from delayed paths leaving the mutual interference among the multiple input symbols

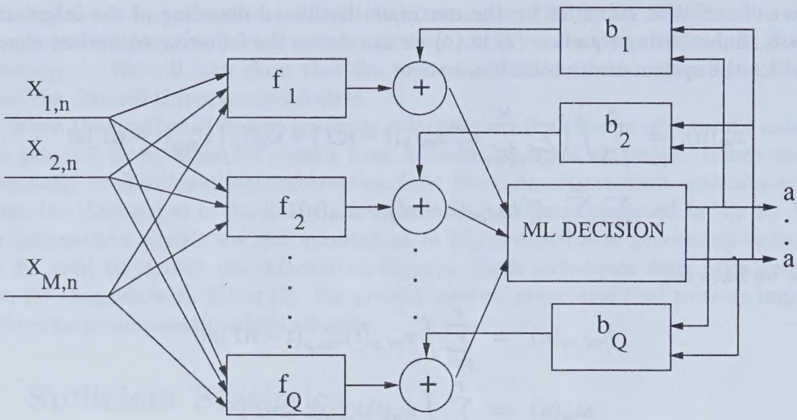


Figure 1: Joint Decision Feedback Equalization

from the respective dominant paths to be addressed by the maximum likelihood space-time decoder. This will guarantee that the full diversity that the codes are designed for is indeed achieved. This equalization approach can also be used for BLAST/MIMO systems with small number of inputs where maximum likelihood demodulation of the multiple inputs can be done after ISI cancellation. This approach is illustrated in Figure 1.

4.1 Linear Equalization

A linear equalizer receiver structure is illustrated in Figure 2. The sufficient statistics are filtered through linear filters that suppress the ISI. The filtering is followed by maximum likelihood space-time decoding as in Section 2 for space-time codes using the filters outputs as the received signal in the maximum likelihood metric calculation (1). The goal of the filters is to suppress the ISI from the multiple input signals corresponding to the multiple transmit antennas while maximizing the signal component from all the transmit antennas. Note that we do not want to suppress the mutual interference among the different input signals. The choice of the number of receive filters Q to be used depends on the number of transmit and receive antennas M, P and the desired level of complexity of the ML processing.

We present the development for obtaining the filters assuming that the filters and channel impulse response are infinite length and are in the space of square summable sequences l_2. Let the Q filters be represented by $\mathbf{f}^q, 1 \leq q \leq Q$. Each filter is itself composed of M sub-filters one for each sufficient statistic as shown in Figure 2. Denote by $f_m^q(k)$ the k^{th} coefficient of m^{th} sub-filter of the q^{th} filter. Then output of the q^{th} filter is given by

$$y(n) = \sum_{m=1}^{M} \sum_{k} f_m(k) x_m(n-k). \tag{9}$$

We introduce vector-matrix notations to represent the filtering. For each $\{h_{m',m}(k)\}$ define the associated vector $\mathbf{h}_{m',m} \in l_2$ with $(\mathbf{h}_{m',m})_k = h_{m',m}(-k)$ for $k = 0, \pm 1, \pm 2 \cdots ....$ Furthermore, define the time-shifted vectors $\mathbf{h}_{m',m}(n) \in l_2$ with $(\mathbf{h}_{m',m}(n))_k = h_{m',m}(n-$

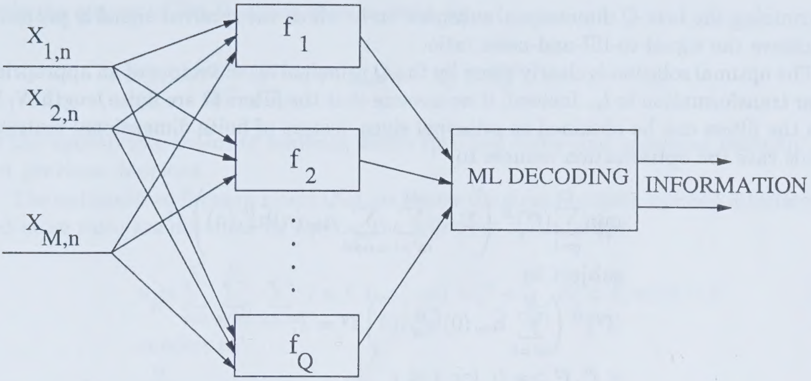


Figure 2: Linear Equalization

k). Then define the vectors $\mathbf{h}_{m'}(n) = [\mathbf{h}^t_{m',1}(n) \cdots \mathbf{h}^t_{m',M}(n)]^t$

Similarly define the noise vector $\mathbf{w}_m \in l_2$ with $(\mathbf{w}_m)_k = w_m(-k)$ for $k = 0, \pm 1, \pm 2 \cdots$.... Then define the time-shifted vectors $\mathbf{w}_m(n) \in l_2$ with $(\mathbf{w}_m(n))_k = w_m(n-k)$ and the vectors $\mathbf{w}(n) = [\mathbf{w}^t_1(n) \cdots \mathbf{w}^t_M(n)]^t$. Define the vector $\mathbf{f}^q_m \in l_2$ such that $(\mathbf{f}^q_m)_k = f^q_m(k)$ for $k = 0, \pm 1, \pm 2 \cdots$ and the vector $\mathbf{f}^q$ obtained by stacking the vectors $\mathbf{f}^q_m$: $\mathbf{f}^q = [(\mathbf{f}^q_1)^t \cdots (\mathbf{f}^q_M)^t]^t$. The output of the q^{th} filter at time n is then, using equation (2), given by

$$
\begin{aligned}
y_q(n) &= \sum_{m=1}^{M} \sum_{k} f^q_m(k) \left(\sum_{m'=1}^{M} \sum_{k'} h_{m',m}(n-k-k') a^{m'}_{k'} + w_m(n-k) \right) \\
&= \sum_{m'=1}^{M} \sum_{k'} \left(\sum_{m=1}^{M} \sum_{k} f^q_m(k) h_{m',m}(n-k-k') \right) a^{m'}_{k'} + \sum_{m=1}^{M} \sum_{k} f^q_m(k) w_m(n-k) \\
&= \sum_{m'=1}^{M} \sum_{k'} < \mathbf{f}^q, \mathbf{h}_{m'}(n-k') > a^{m'}_{k'} + < \mathbf{f}^q, \mathbf{w}(n) >
\end{aligned}
$$

We formulate the following optimality criterion that minimizes sum of the ISI and noise power at the output of the Q filters subject to the anchor condition that the signal power is non-zero. Furthermore, we impose the condition that the noise at the output of the Q filters is independent since ML processing that follows the filtering requires that the noise be independent. Equivalently one can view the optimality criterion as one that maximizes the signal-to-ISI-and-noise ratio, where the "signal" is all of the signals from the different transmit antennas. The optimality criterion is given by

$$\min_{\mathbf{f}^q} \sum_{q=1}^{Q} S_a \sum_{m'=1}^{M} \sum_{n, n \neq 0} | < \mathbf{f}^q, \mathbf{h}_{m'}(n) > |^2 + E | < \mathbf{f}^q, \mathbf{w}(0) > |^2 \tag{10}$$

subject to

$$\sum_{m'=1}^{M} | < \mathbf{f}^q, \mathbf{h}_{m'}(0) > |^2 = 1 \text{ for } q = 1, \cdots, Q \tag{11}$$

$$\text{and} \tag{12}$$

$$< \mathbf{f}^i, \mathbf{f}^j > = 0 \text{ for } i \neq j \tag{13}$$

where S_a denotes the signal power and is assumed to be the same for all the M antenna inputs and $< \cdot, \cdot >$ denotes the inner product in l_2. The optimal filters can be viewed as determining the best Q dimensional subspace on to which the received signal is projected to achieve the signal-to-ISI-and-noise ratio.

The optimal solution is clearly given by the Q principal eigen vectors of an appropriate linear transformation in l_2. Instead, if we assume that the filters $\mathbf{f}^q$ are finite length $N_f M$, then the filters can be obtained as principal eigen vectors of finite dimensional matrices. In this case the optimization reduces to

$$\begin{aligned}
&\min_{\mathbf{f}^q} \sum_{q=1}^{Q} (\mathbf{f}^q)^H \left(\boldsymbol{\Sigma} + \sum_{m'=1}^{M} \sum_{n,n\neq 0} \tilde{\mathbf{h}}_{m'}(n)\tilde{\mathbf{h}}_{m'}^H(n) \right) \mathbf{f}^q \\
&\text{subject to} \\
&(\mathbf{f}^q)^H \left(\sum_{m'=1}^{M} \tilde{\mathbf{h}}_{m'}(0)\tilde{\mathbf{h}}_{m'}^H(0) \right) \mathbf{f}^q = 1; \\
&< \mathbf{f}^i, \mathbf{f}^j >= 0 \ \text{ for } \ i \neq j
\end{aligned}$$

where $\tilde{\mathbf{h}}_{m'}(n)$ are length $N_f M$ windowed versions of $\tilde{\mathbf{h}}$ given by

$$\begin{aligned}
\tilde{\mathbf{h}}_{m'}(n) &= [\tilde{h}^t_{m',1}(n) \cdots \tilde{h}^t_{m',M}(n)]^t \\
\tilde{\mathbf{h}}_{m',m}(n) &= [h_{m',m}(n+\Delta) \cdots h_{m',m}(n+\Delta+N_f-1)]^t
\end{aligned} \tag{14}$$

for decision delay Δ. Thus the optimal $\mathbf{f}^q$ are given by the Q principal eigen vectors of the matrix

$$\left(\boldsymbol{\Sigma} + \sum_{m'=1}^{M} \sum_{n,n\neq 0} \tilde{\mathbf{h}}_{m'}(n)\tilde{\mathbf{h}}_{m'}^H(n) \right)^{-1} \left(\sum_{m'=1}^{M} \tilde{\mathbf{h}}_{m'}(0)\tilde{\mathbf{h}}_{m'}^H(0) \right).$$

It is clear that at most $Q \leq M$ filters are required since the dimension of the desired signal subspace, from the anchor condition (11) is at most M. When there are only a small number of receive antennas $P < M$, depending on the multipath the number of eigen vectors with significant eigen values will be between P and M. Hence fewer than M filters will be sufficient. Note that the ML metric calculation in the space-time decoder following the filtering can then use the Q filter outputs as though they were equivalent to Q received antennas in (1). Setting $Q = 1$ for large P so as to reduce the complexity of the ML metric calculation can be viewed as the generalization of the idea of PRC [9] to frequency-selective fading channels. Furthermore, it is clear that one can tradeoff performance for reduced complexity at various levels by appropriate choice of Q. PRC is simply the special case of this technique of trading performance for complexity for which $Q = 1$.

4.2 Decision Feedback Equalization

The structure of the decision feedback equalization analogous to the MMSE-DFE for SISO systems is shown in Figure 1. The primary application for the DFE would be for the BLAST/MIMO system. For the trellis space-time coded system the feedback taps can be used in the Viterbi decoder as in a parallel decision feedback decoding for SISO systems [3]. As in the linear equalization case the sufficient statistics are filtered by Q feedforward filters $\mathbf{f}_q, 1 \leq q \leq Q$. The Q inputs to ML decision device assuming perfect decisions of past symbols is given in the vector-notation by

$$y^q(n) = \sum_{m'=1}^{M} \sum_{k' \geq n} < \mathbf{f}^q, \mathbf{h}_{m'}(n-k') > a^{m'}_{k'} + < \mathbf{f}^q, \mathbf{w}(n) >, \ \ q = 1, \cdots, Q$$

while the output of the feedback filter is given by

$$z^q(n) = \sum_{m'=1}^{M} \sum_{k'<n} < \mathbf{f}^q, \mathbf{h}_{m'}(n-k') > a_{k'}^{m'}, \quad q = 1, \cdots, Q$$

for the appropriate choice of feedback filters specified below and assuming perfectly correct previous decisions.

The optimal feed forward filters that maximize the signal-to-inter-symbol-interference-and-noise ratio are obtained by solving the following optimization:

$$\min_{\mathbf{f}^q} \sum_{q=1}^{Q} \sum_{m'=1}^{M} \sum_{n,n>0} | < \mathbf{f}, \mathbf{h}_{m'}(-n) > |^2 + \sum_{q=1}^{Q} E| < \mathbf{f}, \mathbf{w}(0) > |^2 \tag{15}$$

subject to

$$\sum_{m'=1}^{M} | < \mathbf{f}^q, \mathbf{h}_{m'}(0) > |^2 = 1, q = 1, \cdots, Q$$

$$< \mathbf{f}^i, \mathbf{f}^j >= 0, \quad i \neq j, 1 \leq i, j \leq Q.$$

The optimality criterion above differs from that of the linear equalizer in that in (15) only the energy in the precursor ISI terms have to be minimized while the postcursor ISI is cancelled by the decision feedback section of the equalizer.

For finite filter lengths, the optimal solution $\mathbf{f}^*$ is given by the Q principal eigen vectors of the matrix

$$\left(\Sigma + \sum_{m'=1}^{M} \sum_{n,n>0} \tilde{\mathbf{h}}_{m'}(-n) \tilde{\mathbf{h}}_{m'}^H(-n) \right)^{-1} \left(\sum_{m'=1}^{M} \tilde{\mathbf{h}}_{m'}(0) \tilde{\mathbf{h}}_{m'}^H(0) \right).$$

where the $\tilde{\mathbf{h}}_m(n)$ are defined as in equation (14). The feedback filter coefficients are given by

$$(\mathbf{b}_{m'}^q)_n \triangleq (\mathbf{f}^q)^H \mathbf{h}_{m'}(n), \quad n = 0, 1, 2, \cdots$$

5 Simulation Results

The performance of the 4-state QPSK code that achieves 2 bit/sym throughput for a channel with two independent Rayleigh fading paths with the delayed path 15 dB average power below the main path was simulated. The second path was assumed to be delayed by 1 symbol relative to the main path. Figure 3 shows the diversity gain that is achieved. The 1X1 system also had the same power profile. The Figure also shows the performance without an equalizer showing the importance of an equalizer for this channel. The PER plots are obtained by averaging over 1000 instantiations of the multipath channel with a different channel instance for each packet.

Figure 3 on the right shows the improvement that can be achieved by doing ML processing after linear equalization as in Figure 2 over simple linear processing to separate the two independent streams for the 2X2 case. In this simulation, BPSK data was transmitted out of each antenna with no coding. With larger number of antennas and larger constellations it may not be practical to implement ML demodulation for the multiple inputs. Nevertheless, the linear equalization can be used in combination with some suboptimum but joint demodulation such as successive interference cancellation [6].

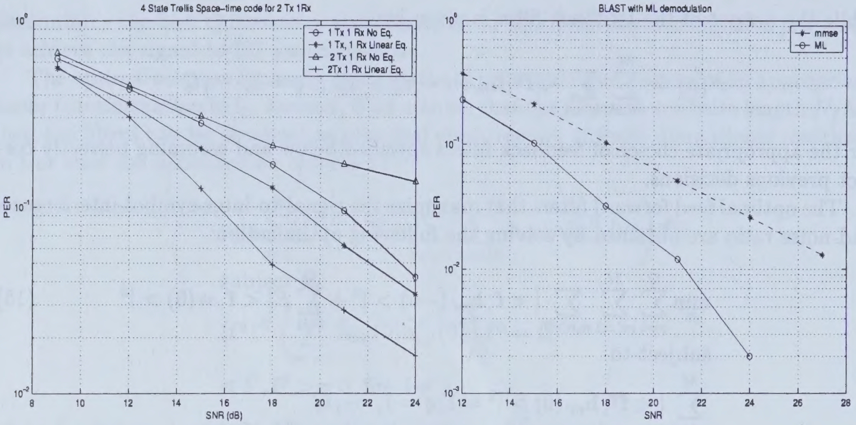


Figure 3: Linear Equalization

Figure 4 shows the performance improvement when the decision feedback equalizer method described in Section 5.2 is used to determine the feedforward and feedback filters with feedback filters implemented within the space-time decoder on a per path basis although the input to the Viterbi decoder is not white as in the case of a traditional PDFD decoder [3].

6 Summary and Future Work

Equalization for space-time coded multi-antenna systems was considered. Linear and DFE equalizers were derived based on a new optimality criterion that essentially maximizes the joint signal energy while suppressing the inter-symbol interference. A new algorithm for improving the performance of BLAST/MIMO for small number of antennas was given based on ML processing at the receiver. This method of front-end equalization also allows the use of other techniques such as successive cancellation [6] that were derived for flat-fading channels. Simulation results showing gains from using the various equalization schemes derived here were presented.

References

[1] S. Al Alamouti, "A simple scheme for diversity transmission", *IEEE Journal on Selected Areas in Communications*, vol. 16, pp. 1451-1458, Oct. 1998.

[2] S.L. Ariyavisitakul, J.H. Winters, and I. Lee, "Optimum space-time processors with dispersive interference: unified analysis and required filter span", *IEEE Transactions on Communications*, vol. 47, no. 7, pp. 1073-1083, July, 1999.

[3] M.V. Eyuboglu and S.U. Qureshi, "Reduced-state sequence estimation with set partitioning and decision feedback", *IEEE Transactions on Communications*, vol. COM-36, pp. 13-26, Sept. 1987.

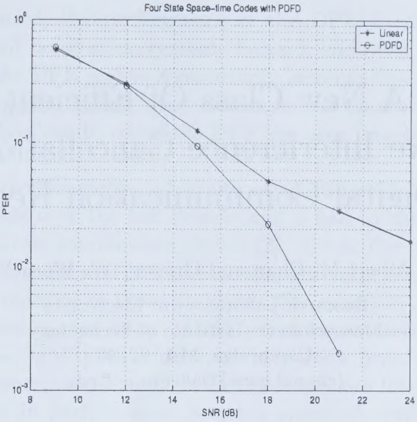


Figure 4: Parallel Decision Feedback Decoding for Space-time Codes

[4] G. J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multiple antennas", *Bell Labs Technical Journal*, vol. 1, no. 1, pp. 41-59, Autumn 1996.

[5] G.J. Foschini and M. J. Gans, "On limits of wireless communication in a fading environment when using multiple antennas", *Wireless Personal Communications*, vol. 6, no. 3, pp. 311-335, March 1998.

[6] G. J. Foschini, G.D. Golden, R.A. Valenzuela, and P.W. Woliansky "Simplified processing for high spectral efficiency wireless employing multi-element antenna arrays", *IEEE Journal on Selected Areas on Communications*, vol. 17, no. 11, November 1999, pp. 1841-52.

[7] A. Lozano, C. Papadias, G.J. Foschini, and R.A. Valenzuela, "An optimal space-time BLAST receiver for wideband wireless communications", *Bell Labs Technical Memorandum*, BL01131K-99-08-31-7TM, August 1998.

[8] A.F. Naguib, V. Tarokh, N. Seshadri, and A.R. Calderbank, "Combined array processing and space-time coding", *IEEE Transactions on Information Theory*, vol. 45, no. 4, pp. 1121-1128, May 1989.

[9] V. Tarokh and T.K.Y. Lo, "Prinicipal ratio combining for fixed wireless applications when transmitter diversity is employed", *IEEE Communications Letters*, vol. 2, no. 8, August 1998.

[10] V. Tarokh, N. Seshadri and A. R. Calderbank, "Space-time codes for high data rate wireless communications", *IEEE Transactions on Information Theory*, vol. 44, no.2, pp. 744-765, March 1998.

[11] H. Viswanathan, "Equalization of Space-time Coded Multi-antenna Systems", submitted to the *IEEE Transactions on Communications*, October, 2000

A New Class Of Efficient Block-Iterative Interference Cancellation Techniques For Digital Communication Receivers

Albert M. Chan and Gregory W. Wornell
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139
{chanal,gww}@allegro.mit.edu

Abstract

A new and efficient class of nonlinear receivers is introduced for digital communication systems. These "iterated-decision" receivers use optimized multipass algorithms to successively cancel interference from a block of received data and generate symbol decisions whose reliability increases monotonically with each iteration. Two variants of such receivers are discussed: the iterated-decision equalizer and the iterated-decision multiuser detector. Iterated-decision equalizers, designed to equalize intersymbol interference (ISI) channels, asymptotically achieve the performance of maximum-likelihood sequence detection (MLSD), but only have a computational complexity on the order of a linear equalizer (LE). Even more importantly, unlike the decision-feedback equalizer (DFE), iterated-decision equalizers can be readily used in conjunction with error-control coding. Iterated-decision multiuser detectors, designed to cancel multiple-access interference (MAI) in CDMA channels, asymptotically achieve the performance of the optimum multiuser detector in uncoded systems, but only have a computational complexity on the order of a decorrelating detector or a linear minimum mean-square error (MMSE) multiuser detector.

1 Introduction

Over the last several decades, a variety of equalization techniques have been proposed for use on intersymbol interference (ISI) channels. Linear equalizers (LE) are attractive from a complexity perspective, but often suffer from excessive noise enhancement. Maximum-likelihood sequence detection (MLSD) is an asymptotically optimum receiver in terms of bit-error rate performance, but its high complexity has invariably precluded its use in practice. Decision-feedback equalizers (DFE) [1] are a widely used compromise, retaining a complexity comparable to the LE, but incurring much less noise enhancement. However, DFEs still have some serious shortcomings. First, decisions made at the slicer can only be fed back to improve future decisions due to the sequential way in which the receiver processes data. Thus, only postcursor ISI can be subtracted, so even if ideal postcursor

---

This work has been supported in part by NSF under Grant No. CCR-9979363, Qualcomm, Inc, and Sanders, a Lockheed-Martin Company.

ISI cancellation is assumed, the performance of the DFE is still limited by possible residual precursor ISI and noise enhancement. Second, and even more importantly, the sequential structure of the DFE makes it essentially incompatible for use in conjunction with error-control coding (on channels not known at the transmitter, as is the case of interest in this paper). As a result, use of the DFE has been largely restricted to uncoded systems.

In parallel to these developments, a variety of multiuser detectors have been proposed for CDMA channels over the last decade and a half as solutions to the problem of mitigating multiple-access interference (MAI) [2]. Given the close coupling between the problems of suppressing ISI and MAI, there are, not surprisingly, close relationships between corresponding solutions to these problems. For example, decorrelating detectors and linear minimum mean-square error (MMSE) multiuser detectors—the counterparts of zero-forcing and linear MMSE equalizers are attractive from a complexity perspective, but suffer from noise enhancement. Optimum maximum likelihood (ML) multiuser detection, while superior in performance, is not a practical option because of its high complexity.

In this paper, we introduce a class of remarkably efficient multipass receivers that is a particularly attractive alternative to all these conventional equalizers and detectors. Specifically, in Section 2 we describe the iterated-decision equalizer, and in Section 3 we describe the corresponding multiuser iterated-decision detector. In particular, we show that these new receivers achieve asymptotically optimum performance while retaining low complexity.

2 The Iterated-Decision Equalizer

In the discrete-time baseband model of the pulse amplitude modulation (PAM) communication system we consider, the transmitted data is a white M-ary phase-shift keying (PSK) stream of coded or uncoded symbols $x[n]$, each with energy $\mathcal{E}_s$. The symbols $x[n]$ are corrupted by a convolution with the channel impulse response, $a[n]$, and by additive noise, $w[n]$, to produce the received symbols[1] $r[n] = a[n] * x[n] + w[n]$. The noise $w[n]$ is a zero-mean, complex-valued, circularly symmetric, stationary, white Gaussian noise sequence with variance $\mathcal{N}_0$ that is independent of $x[n]$. (To simplify the exposition, we focus on symbol-spaced equalization. However, the fractionally spaced generalizations required in practice follow in a straightforward manner, as developed in [4].)

Also, as increasingly aggressive data rates are pursued in wideband systems to meet escalating traffic requirements, ISI becomes increasingly severe. Accordingly, in this paper we pay special attention to the performance and properties of the equalizers in this regime. For the purposes of analysis, a convenient severe-ISI channel model we will exploit is one in which $a[n]$ is a finite impulse response (FIR) filter of length L, where L is large and the taps are mutually independent, zero-mean, complex-valued, circularly symmetric Gaussian random variables with variance σ_a^2. The channel taps $a[n]$ are also independent of the data $x[n]$ and the noise $w[n]$. Note that this is also a good channel model for many wireless systems employing transmitter antenna diversity in the form of linear space-time coding [5].

We begin by focussing on the case in which the receiver has exact knowledge of $a[n]$ in order to develop the basic theory and fundamental limits. In Section 2.1, we consider

---

[1] The symbol $*$ denotes the convolution operation.

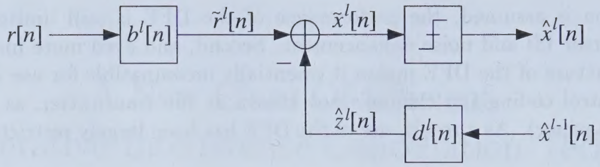


Figure 1: Iterated-decision equalizer structure.

the case of practical interest in which the receiver does not have *a priori* knowledge of the channel, and develop a corresponding adaptive equalizer. We emphasize that in both cases, we restrict our attention to transmitters that have no knowledge of the channel, which is the usual case for reasonably rapidly time-varying channels.

The iterated-decision equalizer we now develop processes the received data in a block-iterative fashion [4, 3]. Specifically, during each iteration or "pass," a linear filter is applied to all the received data, and tentative decisions made in the previous iteration are then used to construct and subtract out an estimate of the ISI. The resulting ISI-reduced data is then passed on to a slicer, which makes a new set of tentative decisions. With each successive iteration, increasingly refined hard decisions are generated using this strategy.

The detailed structure of the iterated-decision equalizer is depicted in Fig. 1.[2]

The parameters of all systems and signals associated with the lth pass are denoted using the superscript l. On the lth pass of the equalizer where $l = 1, 2, 3, \ldots$, the received data $r[n]$ is first processed by a linear filter $b^l[n]$, producing the sequence $\tilde{r}^l[n] = b^l[n] * r[n]$. Next, an appropriately constructed estimate $\hat{z}^l[n]$ of the ISI is subtracted from $\tilde{r}^l[n]$ to produce $\tilde{x}^l[n]$, i.e., $\tilde{x}^l[n] = \tilde{r}^l[n] - \hat{z}^l[n]$, where $\hat{z}^l[n] = d^l[n] * \hat{x}^{l-1}[n]$. Since $\hat{z}^l[n]$ is intended to be some kind of ISI estimate, we impose the constraint that the zero-delay tap of $d^l[n]$ be zero or, equivalently, that $\frac{1}{2\pi}\int_{-\pi}^{\pi} D^l(\omega)d\omega = 0$. The slicer then generates the hard decisions $\hat{x}^l[n]$ from $\tilde{x}^l[n]$ using a minimum-distance rule.

Analogous to counterparts in [5] and [6], we have that when $x[n]$ and $\hat{x}^{l-1}[n]$ are sequences of zero-mean uncorrelated symbols, each with energy $\mathcal{E}_s$, such that their normalized correlation is of the form $E[x^*[n] \cdot \hat{x}^{l-1}[k]]/\mathcal{E}_s \approx \rho^{l-1}\delta[n-k]$, then the slicer input at each iteration can be expressed as

$$\tilde{x}^l[n] \approx E[AB^l]x[n] + v^l[n] \tag{1}$$

where $A(\omega)$ and $B^l(\omega)$ are the frequency responses of $a[n]$ and $b^l[n]$ respectively, where $v^l[n]$ is a complex-valued, marginally Gaussian, zero-mean white noise sequence, uncorrelated with the input symbol stream $x[n]$ and whose variance is a function of $B^l(\omega)$ and $D^l(\omega)$, and where the accuracy of the approximation in (1) increases with the length L of the impulse response $a[n]$.

The second-order model (1) turns out to be a useful one for analyzing and optimizing the performance of the iterated-decision equalizer. In particular, the signal-to-interference+noise ratio (SINR) at the slicer input during each pass, defined as $\gamma^l =$

---

[2] As developed in [3, 4], for coded systems, it suffices to replace the slicer in Fig. 1 with the associated decoder, and to re-encode the previous iteration's decisions $\hat{x}^{l-1}[n]$ before they are input to $d^l[n]$. Note that for DFE's, this kind of integration of decoding with equalization is not possible due to causality constraints.

$\mathcal{E}_s|E[AB^l]|^2/\mathrm{var}\ v^l[n]$, achieves a maximum value of

$$\gamma^l = \left(\frac{1}{1 - \mathcal{E}_s(1-(\rho^{l-1})^2) E\left[\frac{|A|^2}{\mathcal{N}_0 + \mathcal{E}_s(1-(\rho^{l-1})^2)|A|^2}\right]} - 1 \right) \cdot \frac{1}{1-(\rho^{l-1})^2} \tag{2}$$

when

$$\begin{aligned} B^l(\omega) &\propto \frac{A^*(\omega)}{\mathcal{N}_0 + \mathcal{E}_s(1-(\rho^{l-1})^2)|A(\omega)|^2} && (3) \\ D^l(\omega) &= \rho^{l-1}\left(A(\omega)B^l(\omega) - E[AB^l]\right). && (4) \end{aligned}$$

While the strictly causal feedback filter of the DFE subtracts out only postcursor ISI, the noncausal nature of the optimal $D^l(\omega)$ allows the iterated-decision equalizer to cancel both precursor and postcursor ISI. Moreover, the expression for the optimal $D^l(\omega)$ is intuitively satisfying. If $\hat{x}^{l-1}[n] = x[n]$ so that $\rho^{l-1} = 1$, then the output of $D^l(\omega)$ exactly reproduces the ISI component of $\tilde{r}^l[n]$. More generally though, the correlation coefficient ρ^{l-1} describes our confidence in the quality of the estimate $\hat{x}^{l-1}[n]$ and scales the ISI estimate accordingly. Note that the center tap of $d^l[n]$ is indeed zero, as stipulated earlier. Note also that during the first ($l = 1$) pass, the feedback branch is not used because $\rho^0 = 0$, so the sequence $\hat{x}^0[n]$ does not need to be defined.

Next, the properties of $v^l[n]$ imply that the probability of symbol error at the lth iteration is well-approximated by the high signal-to-noise ratio (SNR) formula for the M-ary PSK symbol error rate of a symbol-by-symbol threshold detector for additive white Gaussian noise (AWGN) channels, given by [7]

$$\Pr(\epsilon^l) = 2\mathcal{Q}\left(\sin\left(\frac{\pi}{M}\right)\sqrt{2\gamma^l}\right), \tag{5}$$

where $\mathcal{Q}(v) = \frac{1}{\sqrt{2\pi}}\int_v^\infty e^{-t^2/2}dt$.

By substituting (3) and (4) into (2) and exploiting the fact that arbitrarily close samples of the random process $A(\omega)$ are asymptotically independent in the limit as $L \to \infty$, it can be shown that

$$\gamma^l = \left(\frac{1}{\xi^l e^{\xi^l} E_1(\xi^l)} - 1\right) \cdot \frac{1}{1-(\rho^{l-1})^2} \tag{6}$$

where $E_1(s) = \int_s^\infty e^{-t}/t\ dt$ and $\xi^l = \zeta/(1-(\rho^{l-1})^2)$, with $1/\zeta = \mathcal{E}_s L\sigma_a^2/\mathcal{N}_0$ being the expected SNR at which the transmission is received.

As the above development suggests, successful implementation of the iterated-decision equalizer requires that the normalized correlation between the symbols $x[n]$ and $\hat{x}^{l-1}[n]$ at the lth pass be computable. The iterative algorithm for computing the set of correlation coefficients ρ^l, and in turn predicting the sequence of symbol error probabilities is as follows.

1. Set $\rho^0 = 0$ and let $l = 1$.
2. Compute the SINR γ^l at the slicer input on the lth decoding pass from ρ^{l-1} via (6). [It is worth pointing out that for shorter ISI channels, we can alternatively (and in some cases more accurately) compute γ^l from ρ^{l-1} via (2), where the expectation is replaced by a frequency average.]
3. Compute the symbol error probability $\Pr(\epsilon^l)$ at the slicer output from γ^l via (5).

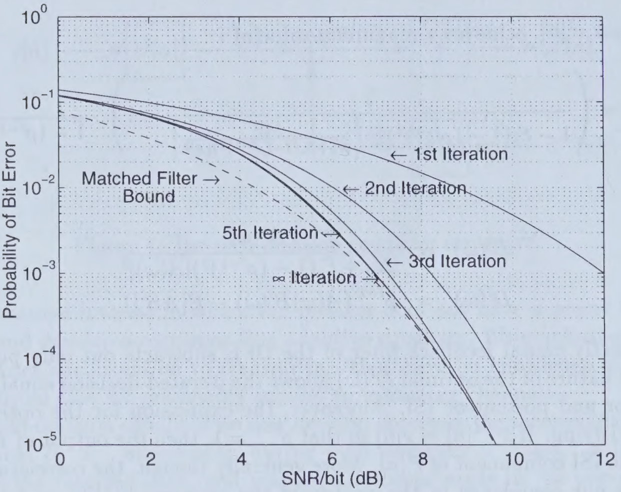


Figure 2: Theoretical QPSK bit-error rate for the iterated-decision equalizer as a function of SNR per bit and the number of decoding iterations.

4. Compute the normalized correlation coefficient ρ^l between the symbols $x[n]$ and the decisions $\hat{x}^l[n]$ generated at the slicer via [6]

$$\rho^l = 1 - 2\sin^2\left(\frac{\pi}{M}\right)\Pr(\epsilon^l). \tag{7}$$

5. Increment l and go to step 2.

A straightforward analysis of the algorithm reveals that the sequence of error probabilities $\Pr(\epsilon^1), \Pr(\epsilon^2), \ldots$ is monotonically decreasing, suggesting that additional iterations always improve performance. However, the error rate performance for a given SNR of $1/\zeta$ eventually converges to a steady-state value of $\Pr(\epsilon^\infty)$.

In Fig. 2, bit error rate is plotted as a function of SNR for $1, 2, 3, 5$, and an infinite number of iterations. We observe that steady-state performance is approximately achieved with comparatively few iterations, after which additional iterations provide only negligibly small gains in performance. It is significant that few passes are required to converge to typical target bit-error rates, since the amount of computation is directly proportional to the number of passes required. We emphasize that the complexity of the iterated-decision equalizer is comparable to that of the DFE or the LE. It is also significant to note that as the SNR increases ($\zeta \to 0$), the slicer input SINR $\gamma \to 1/\zeta$ which, when substituted into (5), corresponds to the matched filter bound for the channel. Thus, perfect ISI cancellation is approached at high SNR.

We plot in Fig. 3 the theoretical performance of the iterated-decision equalizer and the ideal minimum mean-square error decision-feedback equalizer (MMSE-DFE) as $L \to \infty$ [4], as well as the associated experimental performance when $L = 256$. We can readily see that at moderate to high SNR, the iterated-decision equalizer requires significantly less transmit power than the MMSE-DFE to achieve the same error rate. In fact, at high SNR, the iterated-decision equalizer theoretically requires $10\Gamma_0 \log e \approx 2.507$ dB

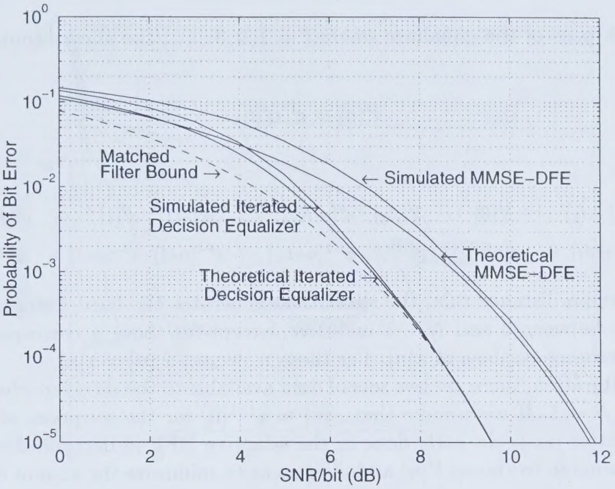


Figure 3: Theoretical and experimentally observed QPSK bit-error rate for the iterated-decision equalizer and the MMSE-DFE as a function of SNR per bit.

less transmit power [4] to achieve the same probability of error as the ideal MMSE-DFE, where $\Gamma_0 = 0.57721\cdots$ denotes Euler's constant.

2.1 Adaptive Implementations

We now develop an adaptive implementation of the iterated-decision equalizer, in which optimal FIR filter coefficients are selected automatically (from the received data) without explicit knowledge of the channel characteristics. We focus on the single channel case; multichannel generalizations follow in a straightforward manner, as developed in [4].

The iterated-decision equalizer is designed to process received data in a block-iterative fashion, so it is ideally suited for packet communication in which the packet size is chosen small enough that the channel encountered by each packet appears linear time-invariant. As is typically the case with other adaptive equalizers, the adaptive iterated-decision equalizer makes use of training symbols sent along in the packet with the data symbols. Suppose that a block of white M-ary PSK symbols $x[n]$ for $n = 0, 1, \ldots, N-1$ is transmitted; some of the symbols (not necessarily at the head of the packet) are for training, while the rest are data symbols.

In the adaptive implementation of the iterated-decision equalizer, the filters $b^l[n]$ and $d^l[n]$ for the lth iteration are finite-length filters. Specifically, $b^l[n]$ has J_1 strictly anticausal taps and J_2 strictly causal taps plus a center tap, while $d^l[n]$ has K_1 strictly anticausal taps and K_2 strictly causal taps with no center tap.

Before the first pass ($l = 1$), we need to initialize the hard decisions $\hat{x}^0[n]$. Since the locations and values of the training symbols in $x[n]$ are known at the receiver, we set $\hat{x}^0[n] = x[n]$ for the n corresponding to those locations. For all the other n between 0 and $N-1$ inclusive, we set $\hat{x}^0[n]$ to be a "neutral" value—for white PSK symbols, this value should be zero.

On the lth pass of the equalizer where $l = 1, 2, 3, \ldots$, the slicer input $\tilde{x}^l[n]$ can be expressed as[3]

$$\tilde{x}^l[n] = \mathbf{c}^{l^\dagger} \mathbf{q}^l[n] \tag{8}$$

where

$$\mathbf{c}^{l^\dagger} = \begin{bmatrix} b^l[-J_1] & \cdots & b^l[0] & \cdots & b^l[J_2] & d^l[-K_1] & \cdots & d^l[-1] & d^l[1] & \cdots & d^l[K_2] \end{bmatrix} \tag{9}$$

$$\mathbf{q}^l[n] = \begin{bmatrix} r[n{+}J_1] & \cdots & r[n] & \cdots & r[n{-}J_2] & \hat{x}^{l-1}[n{+}K_1] & \cdots & \hat{x}^{l-1}[n{+}1] & \hat{x}^{l-1}[n{-}1] & \cdots & \hat{x}^{l-1}[n{-}K_2] \end{bmatrix}^T. \tag{10}$$

Using a minimum-distance rule, the slicer then generates the hard decisions $\hat{x}^l[n]$ from $\tilde{x}^l[n]$ for all n between 0 and $N-1$ inclusive, except for those n corresponding to the locations of training symbols in $x[n]$. For those n, we set $\hat{x}^l[n] = x[n]$.

In the lth iteration, there are two sets of data available to the receiver: $r[n]$ and $\hat{x}^{l-1}[n]$, $n = 0, 1, \ldots, N-1$. If we assume that $x[n] \approx \hat{x}^{l-1}[n]$ for the purposes of determining the optimal filters (as is similarly done in the adaptive DFE in decision-directed mode), then it is reasonable to choose $b^l[n]$ and $d^l[n]$ so as to minimize the sum of error squares:

$$\mathcal{E}(\mathbf{c}^l) = \sum_{n=-\infty}^{\infty} |\hat{x}^{l-1}[n] - \mathbf{c}^{l^\dagger} \mathbf{q}^l[n]|^2. \tag{11}$$

Since this is a linear least-squares estimation problem, the optimum $\mathbf{c}^l$ is [8]

$$\mathbf{c}^l_{\text{opt}} = [\mathbf{\Phi}^l]^{-1} \mathbf{u}^l, \tag{12}$$

where $\mathbf{\Phi}^l = \sum_{n=-\infty}^{\infty} \mathbf{q}^l[n]\mathbf{q}^{l^\dagger}[n]$ and $\mathbf{u}^l = \sum_{n=-\infty}^{\infty} \hat{x}^{l-1^*}[n]\mathbf{q}^l[n]$. The resulting equalizer lends itself readily to practical implementation, even for large filter lengths. In particular, the matrix $\mathbf{\Phi}^l$ can be efficiently computed using correlation functions involving $r[n]$ and $\hat{x}^{l-1}[n]$, and $[\mathbf{\Phi}^l]^{-1}$ can be efficiently computed using formulas for the inversion of a partitioned matrix [9].

We now turn to a couple of implementation issues. First, we would ideally like our finite-length adaptive filters to approximate (3) and (4), which are infinite length. The optimal $b^l[n]$ in (3) includes a filter matched to $a[n]$, and the optimal $d^l[n]$ in (4) includes a cascade of $a[n]$ and the corresponding matched filter, suggesting that a reasonable rule of thumb is to select $J_1 = J_2 = K_1 = K_2 = L$. Second, the block-iterative nature of the equalizer allows the training symbols to be located anywhere in the packet. Since—in contrast to the DFE—the locations do not appear to affect equalizer performance, we arbitrarily choose to uniformly space the training symbols within each packet.

In Fig. 4, we plot the bit-error rate of the adaptive iterated-decision equalizer as a function of the number of iterations, for varying amounts of training data. The graph strongly suggests that there is a threshold for the number of training symbols, below which the adaptive equalizer performs poorly and above which the bit-error rate consistently converges to approximately the same steady-state value regardless of the exact number of training symbols. The excess training data is still important though, since the bit-error rate converges quicker with more training data.

We next examine the probability of bit error as a function of SNR for varying amounts of training data. From Fig. 5 we see that, as expected, performance improves as the amount of training data is increased. Moreover, only a modest amount of training symbols

---

[3] The superscripts ^{T} and † denote the transpose and conjugate-transpose operations, respectively.

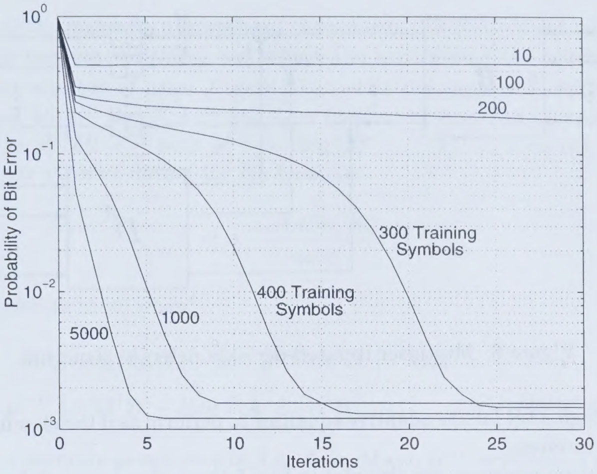


Figure 4: Experimentally observed QPSK bit-error rate for the adaptive iterated-decision equalizer as a function of the number of decoding iterations and the number of training symbols transmitted with each block of 10000 data symbols at an SNR per bit of 7 dB. The 100-tap channels were equalized using 201 feedforward taps and 200 feedback taps.

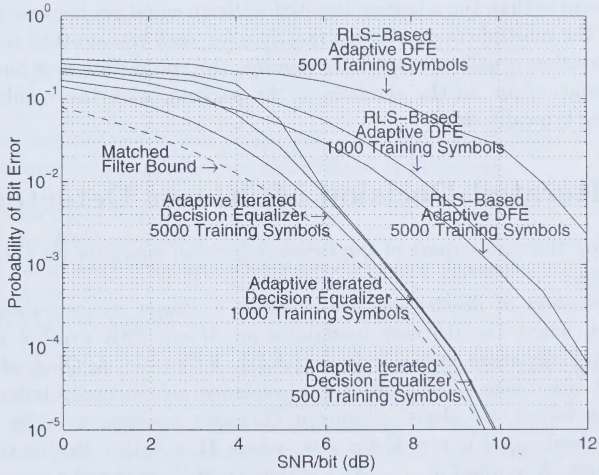


Figure 5: Experimentally observed QPSK bit-error rate for the adaptive iterated-decision equalizer and the RLS-based adaptive DFE (with forgetting factor $\lambda = 1$) as a function of SNR per bit. Blocks of 10000 data symbols were transmitted through 128-tap channels, which were equalized using 257 feedforward taps and 256 noncausal feedback taps in the case of the iterated-decision equalizer, and using 257 feedforward taps and 128 strictly causal feedback taps in the case of the DFE.

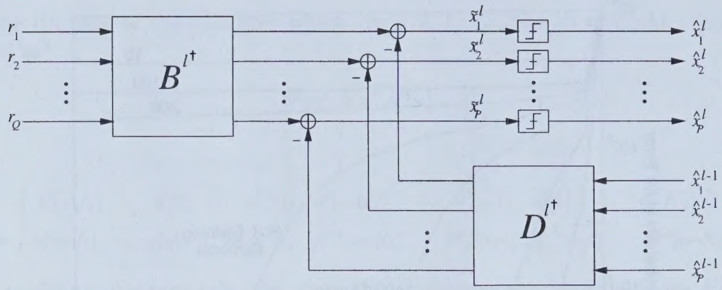


Figure 6: Multiuser iterated-decision detector structure.

is required at high SNR for the adaptive equalizer to perform as if the channel were exactly known at the receiver.

For comparison purposes, we also plot in Fig. 5 the performance of the recursive least squares (RLS) based implementation of the adaptive DFE [8]. The DFE performs significantly worse than the iterated-decision equalizer for comparable amounts of training data. Indeed, the high SNR gap is even larger than the 2.507 dB determined for the non-adaptive case. This is because, as Figs. 3 and 5 show, the performance of the adaptive DFE is not accurately predicted by the nonadaptive MMSE-DFE, even in the long ISI limit. It is also worth stressing that the RLS-based adaptive DFE is much more computationally expensive than the adaptive iterated-decision equalizer because the RLS-based DFE requires the multiplication of large matrices for *each* transmitted symbol, whereas the iterated-decision equalizer essentially requires the computation of one large matrix inverse per iteration for *all* the symbols in the packets, with the number of required iterations being typically small.

3 The Iterated-Decision Multiuser Detector

We now develop the counterpart of the iterated-decision equalizer for the multiuser detection problem.

For the purposes of illustration, we consider a P-user discrete-time synchronous channel model, where the ith user modulates an M-ary PSK symbol x_i onto a randomly generated signature sequence $\mathbf{h}_i = [\ h_i[1] \quad h_i[2] \quad \cdots \quad h_i[Q]\ ]^T$ of length Q assigned to that user, where the taps of the sequence are mutually independent, zero-mean, complex-valued, circularly symmetric Gaussian random variables with variance $1/Q$. The received signal is $\mathbf{r} = \mathbf{HAx} + \mathbf{w}$, where $\mathbf{H} = [\mathbf{h}_1|\cdots|\mathbf{h}_P]$ is the $Q \times P$ matrix of signatures, $\mathbf{A} = \operatorname{diag}\{A_1, \ldots, A_P\}$ is the diagonal matrix of received amplitudes, $\mathbf{x} = [\ x_1 \quad x_2 \quad \cdots \quad x_P\ ]^T$, and $\mathbf{w}$ is a Q-dimensional Gaussian vector with independent zero-mean, complex-valued, circularly symmetric components of variance $\mathcal{N}_0$.

The structure of the multiuser iterated-decision detector is depicted in Fig. 6. The vector of slicer inputs at each iteration is $\tilde{\mathbf{x}}^l = \mathbf{B}^{l\dagger}\mathbf{r} - \mathbf{D}^{l\dagger}\hat{\mathbf{x}}^{l-1}$, where $\mathbf{B}^l = [\mathbf{b}_1|\cdots|\mathbf{b}_P]$, $\mathbf{D}^l = [\mathbf{d}_1|\cdots|\mathbf{d}_P]$, and $\hat{\mathbf{x}}^{l-1} = [\ \hat{x}_1^{l-1} \quad \hat{x}_2^{l-1} \quad \cdots \quad \hat{x}_P^{l-1}\ ]^T$. Analogous to (1) in the case of equalization, the slicer input for the ith user at the lth iteration can be approximated as

$$\tilde{x}_i^l \approx A_i E[\mathbf{b}_i^{l\dagger}\mathbf{h}_i]x_i + v_i^l \tag{13}$$

with $v_1^l, v_2^l, \ldots, v_P^l$ as mutually uncorrelated, zero-mean, complex-valued, circularly symmetric Gaussian random variables, and where the accuracy of the approximation increases when the number of users P gets large while the ratio $\beta \triangleq P/Q$ is held constant. If $\mathbf{I}_P$ and $\mathbf{I}_Q$ are the $P \times P$ and $Q \times Q$ identity matrices respectively, and if $E[x_i^* \cdot \hat{x}_j^{l-1}]/\mathcal{E}_s \approx \rho_i^{l-1}\delta[i-j]$ with $\boldsymbol{\rho}^{l-1} = \operatorname{diag}\{\rho_1^{l-1}, \ldots, \rho_P^{l-1}\}$, then the SINR at the slicer input for the ith user during the lth pass, i.e.,

$$\gamma_i^l = \frac{\mathcal{E}_s A_i E[\mathbf{b}_i^{l\dagger}\mathbf{h}_i]}{\operatorname{var} v_i^l} \tag{14}$$

achieves its maximum value when[4]

$$\begin{aligned} \mathbf{B}^l &\propto \left[\mathcal{N}_0\mathbf{I}_Q + \mathcal{E}_s\mathbf{H}\mathbf{A}(\mathbf{I}_P - \boldsymbol{\rho}^{l-1}\boldsymbol{\rho}^{l-1\dagger})\mathbf{A}^\dagger\mathbf{H}^\dagger\right]^{-1}\mathbf{H}\mathbf{A} && (15) \\ \mathbf{D}^l &= \boldsymbol{\rho}^{l-1\dagger}\left[\mathbf{A}^\dagger\mathbf{H}^\dagger\mathbf{B}^l - \operatorname{diag}\left\{(\mathbf{A}^\dagger E\left[\mathbf{H}^\dagger\mathbf{B}^l\right])_{11}, \ldots, (\mathbf{A}^\dagger E\left[\mathbf{H}^\dagger\mathbf{B}^l\right])_{PP}\right\}\right]. && (16) \end{aligned}$$

In the case of accurate power control, i.e., $\mathbf{A} = A\mathbf{I}_P$ so $\boldsymbol{\rho}^{l-1} = \rho^{l-1}\mathbf{I}_P$, it can be shown that in the limit of large P, the SINR in (14) for each user achieved by these optimum choices of $\mathbf{B}^l$ and $\mathbf{D}^l$ is given by

$$\gamma^l = \left(\frac{1}{1 - \frac{1}{4\beta} \cdot \frac{\mathcal{N}_0}{\mathcal{E}_s(1-\rho^{l-1})^2A^2} \cdot \mathcal{F}\left(\frac{\mathcal{E}_s(1-\rho^{l-1})^2A^2}{\mathcal{N}_0}, \beta\right)} - 1\right) \cdot \frac{1}{1-(\rho^{l-1})^2} \tag{17}$$

where

$$\mathcal{F}(y,z) \triangleq \left(\sqrt{y(1+\sqrt{z})^2+1} - \sqrt{y(1-\sqrt{z})^2+1}\right)^2. \tag{18}$$

The iterative algorithm for computing the correlation coefficients ρ^l, and in turn predicting the sequence of symbol error probabilities is as follows.

1. Set $\rho^0 = 0$ and let $l = 1$.
2. Compute the SINR γ^l from ρ^{l-1} via (17).
3. Compute the symbol error probability $\Pr(\epsilon^l)$ from γ^l via (5).
4. Compute ρ^l via (7).
5. Increment l and go to step 2.

In Fig. 7, we compare the bit-error rate of the multiuser iterated-decision detector with the bit-error rates of various other multiuser detectors as a function of SNR, when decoding $P = 128$ simultaneous users with signatures of length $Q = 128$ with power control. The multiuser iterated-decision detector significantly outperforms the other detectors at moderate to high SNR, and asymptotically approaches the matched filter bound for the single-user channel. Thus, perfect MAI cancellation is approached at high SNR.

---

[4] Using a matrix identity, $\mathbf{B}^l$ can alternatively be computed as

$$\mathbf{B}^l \propto \mathbf{H}\mathbf{A}\left[\mathcal{N}_0\mathbf{I}_P + \mathcal{E}_s(\mathbf{I}_P - \rho^{l-1}\rho^{l-1\dagger})\mathbf{A}^\dagger\mathbf{H}^\dagger\mathbf{H}\mathbf{A}\right]^{-1},$$

which may be easier to evaluate depending on relative sizes of P and Q.

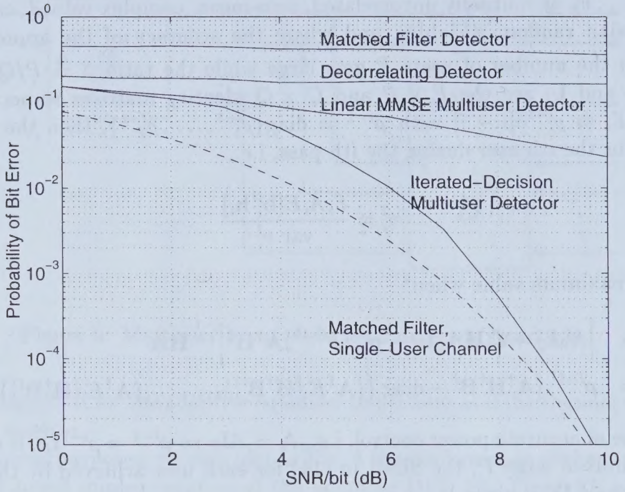


Figure 7: Experimentally observed QPSK bit-error rate for the multiuser iterated-decision detector and various other multiuser detectors as a function of SNR per bit. The number of simultaneous users is 128, and the signature length is 128.

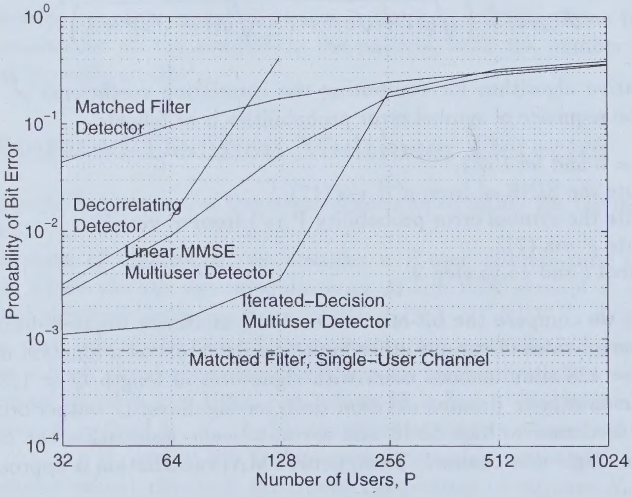


Figure 8: Experimentally observed QPSK bit-error rate for the multiuser iterated-decision detector and various other multiuser detectors as a function of the number of simultaneous users. The signature length is 128, and the SNR per bit is 7 dB.

Fig. 8 depicts the bit-error rates of the detectors as a function of the number of users P with signatures of length $Q = 128$, an SNR per bit of 7 dB, and power control. The multiuser iterated-decision detector has clearly superior performance when $P \lesssim 2Q$.

As a final comment, we note that an adaptive implementation of the iterated-decision multiuser detector is straightforward to derive, and is directly analogous to the adaptive iterated-decision equalizer of Section 2.1. Likewise, for coded systems, an iterated-decision multiuser decoder is readily obtained, and takes a form analogous to the iterated-decision equalizer-decoder structure mentioned in Section 2.

References

[1] C. A. Belfiore and J. H. Park, Jr., "Decision-feedback equalization," *Proc. IEEE*, vol. 67, pp. 1143–1156, Aug. 1979.

[2] S. Verdú, *Multiuser Detection.* Cambridge, U.K.: Cambridge Univ. Press, 1998.

[3] A. M. Chan and G. W. Wornell, "A Class of Block-Iterative Equalizers for Intersymbol Interference Channels," in *Int. Conf. Commun.*, (New Orleans, LA), June 2000.

[4] A. M. Chan, "A class of batch-iterative methods for the equalization of intersymbol interference channels," S. M. thesis, M.I.T., Aug. 1999.

[5] G. W. Wornell and M. D. Trott, "Efficient signal processing techniques for exploiting transmit antenna diversity on fading channels," *IEEE Trans. Signal Processing*, vol. 45, pp. 191–205, Jan. 1997.

[6] S. Beheshti, S. H. Isabelle, and G. W. Wornell, "Joint intersymbol and multiple-access interference suppression algorithms for CDMA systems," *European Trans. Telecomm. & Related Technol.*, vol. 9, pp. 403–18, Sep.–Oct. 1998.

[7] J. G. Proakis, *Digital Communications*, 3rd ed. New York: McGraw-Hill, 1995.

[8] S. Haykin, *Adaptive Filter Theory*, 3rd ed. Englewood Cliffs, NJ: Prentice Hall, 1996.

[9] H. Lütkepohl, *Handbook of Matrices.* Chichester, England: Wiley, 1996.

[10] E. Biglieri, "Ungerboeck codes do not shape the signal power spectrum," *IEEE Trans. Inform. Theory*, vol. 32, pp. 595–596, July 1986.

Combined FIR MIMO Decision Feedback Equalization and Space-Time Block Coding over Multipath Fading Channels

Naofal Al-Dhahir and Ayman F. Naguib
AT&T Shannon Laboratory
Florham Park, NJ 07932

Abstract

A finite–length optimized–delay multi–input multi–output mean–square–error decision–feedback equalizer for space–time block–coded transmission over multipath fading channels is presented. Alamouti's space–time block code with 2 transmit and 2 receive antenna on a typical urban EDGE channel is taken as a case study. We propose a combined equalization and decoding scheme under the constraint of linear processing complexity (no trellis search) at the receiver. Performance comparisons are made with the single–transmit/single–receive antenna case and the case of multi–input multi–output feedforward linear equalization only with no decision feedback.

1 Introduction

Enhanced data rates for GSM evolution (EDGE) standardization efforts aim at providing a smooth transition from the current TDMA second generation digital cellular system (IS–136 and GSM) to third generation systems supporting higher data rates and enhanced services [7].

Space–time block codes (STBC) [6, 8, 9] use multiple transmit antennas to build spatial redundancy in the transmitted signal such that maximum spatial diversity is achieved. For example, with 2 transmit antennas and n_o receive antennas, Alamouti's STBC [6] achieves a spatial diversity order of $2n_o$. It has the added advantages of no bandwidth expansion, not requiring channel knowledge at the transmitter, and simple maximum likelihood decoding at the receiver using only linear processing techniques.

The use of multiple transmit and receive antennas over the EDGE physical layer results in a multi–input multi–output (MIMO) fading multipath channel. STBC's were designed for MIMO flat–fading channels [6, 8, 9]. Therefore, implementing them over EDGE while preserving their linear–decoding–complexity feature requires effective MIMO equalization schemes that satisfy the linear receiver processing complexity requirement of STBC. This precludes the use of trellis–based techniques whose decoding complexity is exponential in the MIMO channel memory [1]. MIMO minimum–mean–square–error linear and decision–feedback equalizers (abbreviated henceforth as MIMO MMSE–LE and

---

[1] Defined as the maximum of the memories of the individual single–input single–output channel impulse responses.

MIMO MMSE–DFE, respectively) were derived in [1, 2] assuming infinite–length unrealizable filters. Recently, they have been extended to the case of finite–length realizable filters with finite decision delay in [3]. As we show in this paper, the highly–dispersive and generally non–minimum–phase characteristics of the EDGE channel make the use of the MIMO MMSE–DFE over the MIMO MMSE–LE advantageous from a performance point of view.

In this paper, we apply the FIR MIMO MMSE–DFE theory developed in [3] to STBC for EDGE in a Typical Urban (TU) environment. In particular, we focus on Alamouti's STBC with 2 transmit and 2 receive antennas and show how to combine it with a two–input two–output MMSE–DFE to combat frequency–selective fading under the constraint of linear–processing–only at the receiver. Monte Carlo simulations are used to evaluate the raw bit error rate as a function of the the number of DFE taps and the operating $\frac{E_b}{N_0}$ level.

2 Model and Assumptions

2.1 Input-Output Model

We consider a two–input two–output multipath fading channel corrupted by additive Gaussian noise. The channel impulse response (CIR) from transmit antenna i to receive antenna j is assumed to be an FIR filter of memory ν.

2.2 Space-Time Block Codes for Flat-Fading Channels

In Alamouti's space–time block encoding scheme [6], the input symbols are grouped in pairs of two where symbols c_k and c_{k+1} are transmitted at time k from antennas 1 and 2, respectively. Then, at time $k+1$, the symbol $-c_{k+1}^*$ (where $(.)^*$ denotes the complex–conjugate transpose) is transmitted from antenna 1 and the symbol c_k^* is transmitted from antenna 2.

For a flat–fading channel that is assumed time–invariant within the burst [2], the noisy received signal at antenna 1 at sampling instants k and $k+1$ can be expressed as follows

$$\begin{bmatrix} r_k^{(1)} \\ r_{k+1}^{*(1)} \end{bmatrix} = \begin{bmatrix} p^{(1,1)} & p^{(1,2)} \\ p^{*(1,2)} & -p^{*(1,1)} \end{bmatrix} \begin{bmatrix} c_k \\ c_{k+1} \end{bmatrix} + \begin{bmatrix} n_k^{(1)} \\ n_{k+1}^{*(1)} \end{bmatrix} , \tag{1}$$

where $p^{(1,1)}$ and $p^{(1,2)}$ are the channel gains from transmit antennas 1 and 2, respectively, to receive antenna 1. In matrix notation, we have

$$\mathbf{r}^{(1)} = \mathbf{P}^{(1)}\mathbf{c} + \mathbf{n}^{(1)} . \tag{2}$$

As shown in [6], since $\mathbf{P}^{(1)}$ is an orthogonal matrix, pre–multiplying (2) by $\mathbf{P}^{*(1)}$ decouples the two components of $\mathbf{r}^{(1)}$ as shown below

$$\tilde{\mathbf{r}}^{(1)} = \mathbf{P}^{*(1)}\mathbf{r}^{(1)} = (|p^{(1,1)}|^2 + |p^{(1,2)}|^2)\mathbf{c} + \tilde{\mathbf{n}}^{(1)} , \tag{3}$$

where $\tilde{\mathbf{n}}^{(1)}$ is still white Gaussian noise. This decoupling simplifies the ML decoding rule to

$$\hat{\mathbf{c}} = argmin_{\hat{\mathbf{c}}}|\tilde{\mathbf{r}}^{(1)} - (|p^{(1,1)}|^2 + |p^{(1,2)}|^2)\hat{\mathbf{c}}|^2 .$$

---

[2] Alamouti's scheme requires only that the channel is time invariant over 2 symbol durations. However, for the EDGE scenarios considered here, the coherence time of the channel is large enough such that the channel can be assumed time invariant over the whole burst.

With 2 receive antennas, it is straightforward to show that the decoding rule becomes [6]

$$\hat{\mathbf{c}} = argmin_{\hat{\mathbf{c}}} \sum_{j=1}^{2} |\tilde{\mathbf{r}}^{(j)} - (|p^{(j,1)}|^2 + |p^{(j,2)}|^2)\hat{\mathbf{c}}|^2 . \tag{4}$$

Therefore, decoding involves simple *linear combining* and *separate* ML decoding of c_k and c_{k+1} which requires, for an M–ary input signal constellation, only $2M$ metric computations instead of M^2 metric computations that would be required without exploiting the orthogonal structure of the STBC.

3 Proposed Receiver Structure

3.1 FIR MIMO MMSE-DFE

The FIR MIMO MMSE–DFE for n_i inputs and n_o outputs consists of $n_i n_o$ FIR feedforward filters and n_i^2 FIR feedback filters whose optimum filter settings were derived in [5] for arbitrary input and noise spatio–temporal auto–correlation matrices. In this subsection, we specialize these results for the case of $n_i = n_o = 2$ and white input [3] and noise (i.e. $\mathbf{R}_{xx} = \mathbf{I}_{2(N_f+\nu)}$ and $\mathbf{R}_{nn} = \sigma_n^2 \mathbf{I}_{2N_f}$). Define the Cholesky factorization

$$\mathbf{I}_{2(N_f+\nu)} + \frac{1}{\sigma_n^2}\mathbf{H}^*\mathbf{H} = \mathbf{L}\mathbf{D}\mathbf{L}^* , \tag{5}$$

where $\mathbf{L}$ is a lower—triangular matrix and $\mathbf{D} = diag(d_0, d_1, \cdots, d_{2(N_f+\nu)})$ is a diagonal matrix. Then, the optimum [4] MIMO feedback matrix $\tilde{\mathbf{B}} \overset{def}{=} \left[\begin{array}{cccc} \mathbf{0}_{n_i \times n_i \Delta} & \mathbf{B}_{opt,0} & \mathbf{B}_{opt,1} \; \cdots & \mathbf{B}_{opt,N_b} \end{array} \right]^t = \left[\begin{array}{cc} \mathbf{0}_{n_i \times n_i \Delta} & \mathbf{B}_{opt} \end{array} \right]^t$ is given by

$$\tilde{\mathbf{B}}_{opt} = \mathbf{L} \left[\begin{array}{cc} \mathbf{e}_{2\Delta_{opt}} & \mathbf{e}_{2\Delta_{opt}+1} \end{array} \right] , \tag{6}$$

where $0 \leq \Delta_{opt} \leq N_f + \nu - 1$ is the optimum decision delay (also known as burst timing) which minimizes the equalization minimum mean square error (MMSE) given by $MMSE = d_{2\Delta_{opt}}^{-1} + d_{2\Delta_{opt}+1}^{-1}$ and $\mathbf{e}_j^*$ is the j^{th} unit row vector whose j^{th} entry is equal to one and all other entries equal to zero. In this paper, we set the number of feedback taps N_b equal to the channel memory ν in which case $\Delta_{opt} = N - 1$.

The corresponding optimum FIR MIMO feedforward filter matrix is given by

$$\begin{aligned} \mathbf{W}_{opt}^* &= \tilde{\mathbf{B}}_{opt}^* \mathbf{H}^* (\mathbf{H}\mathbf{H}^* + \sigma_n^2 \mathbf{I}_{2N_f})^{-1} \\ &= \frac{1}{\sigma_n^2} \left[\begin{array}{c} d_{2\Delta_{opt}}^{-1} \mathbf{e}_{2\Delta_{opt}}^* \\ d_{2\Delta_{opt}+1}^{-1} \mathbf{e}_{2\Delta_{opt}+1}^* \end{array} \right] \mathbf{L}^{-1}\mathbf{H}^* . \end{aligned} \tag{7}$$

---

[3] Space-time block codes impose a non–white spatio–temporal auto-correlation profile on the input sequence. However, the resulting $\mathbf{R}_{xx}$ is non-stationary and data-dependent which would significantly complicate the task of estimating it at the receiver. Therefore, we shall assume that $\mathbf{R}_{xx} = \mathbf{I}_{2(N_f+\nu)}$ to reduce the MIMO DFE coefficient computation complexity. We still account for the input spatio-temporal correlation by exploiting the STBC structure in the decoding.

[4] Assuming that the first non–zero matrix tap of $\tilde{\mathbf{B}}_{opt}$ denoted by $\mathbf{B}_{opt,0}$ is *lower triangular*. This assumption was shown in [5] to result in lower equalization mean square error than the more traditional constraint $\mathbf{B}_{opt,0} = \mathbf{I}_{n_i}$.

The signal at the output of the feedforward equalizer is given by

$$\begin{aligned}
\mathbf{r}_k &\overset{def}{=} \begin{bmatrix} r_k^{(1)} \\ r_k^{(2)} \end{bmatrix} \overset{def}{=} \mathbf{W}^*_{opt}\mathbf{y}_{k+N_f-1:k} \\
&= \mathbf{W}^*_{opt}\mathbf{H}\mathbf{x}_{k+N_f-1:k} + \mathbf{W}^*_{opt}\mathbf{n}_{k+N_f-1:k} \\
&\approx \tilde{\mathbf{B}}^*_{opt}\mathbf{x}_{k+N_f-1:k} + \tilde{\mathbf{n}}_k \\
&= \sum_{i=0}^{N_b} \mathbf{B}^*_{opt,i}\mathbf{x}_{k-i} + \tilde{\mathbf{n}}_k \ . \qquad (8)
\end{aligned}$$

Assuming that $\hat{\mathbf{x}}_{k-i} = \mathbf{x}_{k-i}$, [5] the 4 strictly–causal feedback filters cancel post–cursor intersymbol interference due to $\mathbf{B}_{opt,i}$ (where $1 \leq i \leq N_b$) resulting in the following input to the STBC decoder

$$\begin{aligned}
\mathbf{z}_k &\overset{def}{=} \begin{bmatrix} z_k^{(1)} \\ z_k^{(2)} \end{bmatrix} \overset{def}{=} \mathbf{r}_k - \sum_{i=1}^{N_b} \mathbf{B}^*_{opt,i}\hat{\mathbf{x}}_{k-i} \\
&\approx \mathbf{B}^*_{opt,0}\mathbf{x}_k + \tilde{\mathbf{n}}_k \ . \qquad (9)
\end{aligned}$$

Remarks

- Alamouti's space-time block code imposes a non–white spatio–temporal auto-correlation profile on the input sequence. However, the resulting $\mathbf{R}_{xx}$ is non-stationary and data–dependent which would significantly complicate the task of estimating it at the receiver. Therefore, we shall assume that $\mathbf{R}_{xx} = \mathbf{I}_{2(N_f+\nu)}$ to reduce the MIMO DFE coefficient computation complexity. We still account for the input spatio-temporal correlation by exploiting the STBC structure in the decoding.
- We optimize FIR MIMO MMSE–DFE coefficients assuming that the first non–zero matrix tap of $\tilde{\mathbf{B}}_{opt}$, denoted by $\mathbf{B}_{opt,0}$, is *lower triangular.* This assumption was shown in [5] to result in lower equalization mean square error than the more traditional constraint $\mathbf{B}_{opt,0} = \mathbf{I}_{n_i}$. Physically, the lower–triangular constraint on $\mathbf{B}_{opt,0}$ implies input ordering where lower–indexed inputs are detected first and their decisions are used in making decisions on higher–indexed inputs. Mathematically, under this scheme, to detect $x_k^{(i)}$, we use the decisions made on $x_{k-j}^{(l)}$ where $0 < j \leq N_b$ for $l \geq i$ and $0 \leq j \leq N_b$ for $l < i$.
- In the presence of co–channel interference (CCI), the factored matrix in (5) becomes $\mathbf{I}_{2(N_f+\nu)} + \mathbf{H}^*(\sigma_n^2\mathbf{I}_{2N_f} + \mathbf{R}_{CCI})^{-1}\mathbf{H}$ where $\mathbf{R}_{CCI}$ is the CCI auto–correlation matrix.

3.2 Spatial Noise Decorrelation

The two–dimensional noise vector in (9) is correlated (due to the action of the MIMO feedforward filter) and has a spatial correlation matrix given by

$$\mathbf{R}_{\tilde{n}\tilde{n}} = \mathrm{E}[\tilde{\mathbf{n}}_k\tilde{\mathbf{n}}_k^*] = \sigma_n^2\mathbf{W}^*_{opt}\mathbf{W}_{opt} \ .$$

---

[5] For analysis tractability, we assume correct previous decisions. However, this assumption is relaxed in the simulation results of Section 5 where we take the effects of error propagation into account.

To preserve the simplicity of the ML decoding rule of Alamouti's STBC without incurring performance loss due to this spatial noise correlation, we add a spatial noise decorrelating component to the FIR MIMO MMSE–DFE as described in detail next.

Consider the eigen–decomposition [6]

$$\mathbf{R}_{\tilde{n}\tilde{n}} = \mathbf{U}\mathbf{\Gamma}\mathbf{U}^* \,, \tag{10}$$

where $\mathbf{U}$ is an orthogonal matrix of the eigenvectors and $\mathbf{\Gamma}$ is a diagonal matrix of the eigenvalues of $\mathbf{R}_{\tilde{n}\tilde{n}}$. Define the modified FIR MIMO feedforward and feedback filters as follows

$$\bar{\mathbf{W}}^*_{opt} \overset{def}{=} \mathbf{U}^*\mathbf{W}^*_{opt} \;\text{ and }\; \bar{\mathbf{B}}^*_{opt} \overset{def}{=} \mathbf{U}^*\mathbf{B}^*_{opt} \,. \tag{11}$$

The output of this modified feedforward filter is given by

$$\begin{aligned} \bar{\mathbf{r}}_k &\overset{def}{=} \bar{\mathbf{W}}^*_{opt}\mathbf{y}_{k+N_f-1:k} \\ &= \bar{\mathbf{W}}^*_{opt}\mathbf{H}\mathbf{x}_{k+N_f-1:k} + \bar{\mathbf{W}}^*_{opt}\mathbf{n}_{k+N_f-1:k} \\ &= \mathbf{U}^*\mathbf{W}^*_{opt}\mathbf{H}\mathbf{x}_{k+N_f-1:k} + \bar{\mathbf{n}}_k \\ &\approx \bar{\mathbf{B}}^*_{opt}\mathbf{x}_{k+N_f-1:k} + \bar{\mathbf{n}}_k \\ &= \sum_{i=0}^{N_b} \bar{\mathbf{B}}^*_{opt,i}\mathbf{x}_{k-i} + \bar{\mathbf{n}}_k \,. \end{aligned} \tag{12}$$

Note that Equations (8) and (12) have the same form *except* that the noise term in the latter is spatially uncorrelated. Mathematically,

$$\begin{aligned} \mathbf{R}_{\bar{n}\bar{n}} &= \bar{\mathbf{W}}^*_{opt}\mathbf{R}_{nn}\bar{\mathbf{W}}_{opt} \\ &= \mathbf{U}^*\mathbf{R}_{\tilde{n}\tilde{n}}\mathbf{U} = \mathbf{\Gamma} \,. \end{aligned}$$

With this modified FIR MIMO MMSE–DFE, the input to the ML STBC decoder has the form

$$\bar{\mathbf{z}}_k \overset{def}{=} \bar{\mathbf{r}}_k - \sum_{i=1}^{N_b} \bar{\mathbf{B}}^*_{opt,i}\hat{\mathbf{x}}_{k-i} \tag{13}$$

$$\approx \bar{\mathbf{B}}^*_{opt,0}\mathbf{x}_k + \bar{\mathbf{n}}_k \,. \tag{14}$$

Equation (14) describes a 2×2 (approximately) memoryless MIMO channel with spatially uncorrelated additive Gaussian noise. In practice, the channel seen by the ML BSTC decoder is not exactly memoryless due to imperfect equalization.

Remark

We verified by simulations that implementing the spatial noise decorrelator improves the BER. However, It should be emphasized that it does not affect the equalization MMSE as shown next. With the modified MIMO feedforward and feedback filters in (11), the DFE error sequence is given by

$$\mathbf{e}_{k,new} = \bar{\mathbf{B}}^*_{opt}\mathbf{x}_{k:k-N_b} - \bar{\mathbf{W}}^*_{opt}\mathbf{y}_{k+N_f-1:k} = \mathbf{U}^*\mathbf{e}_k \,.$$

Therefore, the achievable equalization MMSE is equal to

$$\begin{aligned} MMSE_{new} &\overset{def}{=} trace(\mathrm{E}[\mathbf{e}_{k,new}\mathbf{e}^*_{k,new}]) \\ &= trace(\mathbf{U}^*\mathbf{R}_{ee}\mathbf{U}) \\ &= trace(\mathbf{R}_{ee}) \quad : \text{since } trace(\mathbf{AB}) = trace(\mathbf{BA}) \text{ and } \mathbf{U}\mathbf{U}^* = \mathbf{I}_{n_i} \\ &= MMSE \end{aligned}$$

---

[6] With 2 transmit antennas, the Hermitian matrix $\mathbf{R}_{\tilde{n}\tilde{n}}$ is 2×2; hence, its eigenvectors and eigenvalues can be determined in closed–form as a fuction of its entries.

3.3 Combined Equalization and STBC

The FIR MIMO MMSE–DFE attempts to convert the 2×2 dispersive noisy channel into a memoryless channel with spatially uncorrelated noise. More specifically, Equation (14) shows that the 2 signals at the equalizer output approximate the output of a 2×2 memoryless noisy channel with channel gains given by the entries of $\bar{\mathbf{B}}^*_{opt,0}$ and noise variances given by the diagonal entries of $\mathbf{\Gamma}$. Hence, the decoding algorithm for the flat-fading case given in [6] can be applied here **except for the following two important modifications :**

1. The j^{th} term ($j = 1, 2$) in the decoding rule in (4) has to be scaled by $\frac{1}{\mathbf{\Gamma}(j,j)}$ to account for the unequal variances of the two components of $\bar{\mathbf{n}}_k$.

2. The FIR MIMO MMSE–DFE is a *symbol-by-symbol* detector while the STBC decoder operates on consecutive *pairs* of symbols. More specifically, we can not compute the equalizer outputs at time $k+1$, denoted by $\bar{z}^{(1)}_{k+1}$ and $\bar{z}^{(2)}_{k+1}$, unless the interference due to $x^{(j)}_{k-i}$ (for $0 \leq i \leq N_b - 1$ and $j = 1, 2$) is cancelled. Hence, tentative decisions on $x^{(j)}_k$, denoted by $\hat{x}^{(j)}_k$, are needed before STBC decoding can start. We propose the following three additional steps prior to the standard STBC decoding algorithm

   - Compute *tentative* decisions on $x^{(j)}_k$ from $\bar{z}^{(j)}_k$ only by using the ML rule (c.f. Equation (14))

     $$\hat{\mathbf{x}}_k \overset{def}{=} \begin{bmatrix} \hat{x}^{(1)}_k \\ \hat{x}^{(2)}_k \end{bmatrix} = argmin_{\hat{\mathbf{x}}_k} |\bar{\mathbf{z}}_k - \bar{\mathbf{B}}^*_{opt}\hat{\mathbf{x}}_k|^2$$

     $$= \sum_{i=1}^{2}\sum_{j=1}^{2} |\bar{z}^{(i)}_k - \bar{\mathbf{B}}^*_{opt,0}(i,j)\hat{x}^{(j)}_k|^2 ,$$

     where $\bar{\mathbf{B}}^*_{opt,0}(i,j)$ denotes the (i,j) entry of $\bar{\mathbf{B}}^*_{opt,0}$. Note that since $\bar{\mathbf{B}}^*_{opt,0}$ is not an orthogonal matrix [7], this ML decoding rule requires M^2 metric computations.

   - Compute $\bar{z}^{(j)}_{k+1}$ (for $j = 1, 2$) by cancelling interfering effect from $x^{(j)}_k, x^{(j)}_{k-1}, \cdots, x^{(j)}_{k-N_b+1}$ using the FIR MIMO MMSE–DFE structure (c.f. Equation (13)).

   - Use $\bar{z}^{(j)}_k$ and $\bar{z}^{(j)}_{k+1}$ to compute a final decision on $x^{(j)}_k$ and $x^{(j)}_{k+1}$ (and hence final decisions on c_k and c_{k+1} using the standard STBC decoding procedure.

3.4 MIMO MMSE-DFE Complexity

The complexity of the proposed receiver is dominated by the complexity of computing the optimum coefficients of the FIR MIMO MMSE–DFE and applying these coefficients to the received data and previous decisions. The first task is performed once every burst (≈ 0.577 msec for EDGE) while the second task is performed once every symbol ($\approx 3.69\mu$

---

[7] It is possible to make $\bar{\mathbf{B}}^*_{opt,0}$ orthogonal by constraining $\mathbf{B}_{opt,0}$ to be orthogonal (of which the popular constraint $\mathbf{B}_{opt,0} = \mathbf{I}_{n_i}$ is a special case) instead of lower triangular. However, it can be shown [5] that the former constraint results in a higher equalization mean square error than the latter constraint. We also verified by simulations that it results in higher BER

sec for EDGE). We use the number of million instructions per second (abbreviated as MIPS) as our complexity measure. We assume that a single complex multiply and add operation is equivalent to one instruction. In Figure 1, we show the MIPS estimate as a function of number of MIMO MMSE–DFE feedforward taps N_f for the T–spaced and $\frac{T}{2}$–spaced cases, corresponding to $l = 1$ and 2, respectively. We assume $n_i = n_o = 2$ and $N_b = \nu = 3$. It is clear from the Figure that the MIPS estimate is within the processing power of a single modern programmable DSP chip.

4 Simulation Results

Our simulation parameters are typical of the EDGE environment. We assume 8–PSK modulation and a typical urban (TU) wireless channel with a linearized GMSK pulse shape. The overall CIR length is effectively 4 symbol periods; i.e., $\nu = 3$. The bit–to–symbol mapping follows a Gray mapping rule where the binary representations of adjacent 8–PSK symbols differ by only one bit. We compare 3 receiver structures :

1. Single–transmit/single–receive antennas with FIR single–input single–output (SISO) MMSE–DFE [4]. The number of feedback taps N_b is set equal to 3 (same as channel memory ν).

2. Two transmit and two receive antennas with Alamouti's STBC and the FIR MIMO MMSE–DFE of Section 3. Each of the 4 feedback filters has 3 taps.

3. Same as Structure 2 above but with the FIR MIMO MMSE linear equalizer (LE) replacing the DFE, i.e., without the 4 feedback filters.

It is important to emphasize that the total transmitted power for the 3 *scenarios is set to be the same.* In other words, with 2 transmit antennas, each antenna transmits at half the transmit power of the single transmit antenna case. Figure 2 shows raw Bit Error Rate (BER) results from which we conclude

- The FIR MIMO MMSE–DFE significantly outperforms the FIR SISO MMSE–DFE. At BER=10^{-3}, the performance gain with $N_f = 8$ is about 8.5 dB.

- Increasing N_f from 4 to 8 achieves a 2 dB gain at BER=10^{-4}. We found by simulations that increasing N_f beyond 8 results in negligible BER improvement.

- The FIR MIMO MMSE–DFE achieves significantly lower BER than the FIR MIMO MMSE–LE when both structures are constrained to have the same (total) number of taps. For example, at BER=10^{-3}, a MIMO MMSE–DFE with $N_f = 8$ and $N_b = 3$ (i.e. a total of 44 taps) requires 4.5 dB less $\frac{E_b}{N_0}$ than the MIMO MMSE–LE with 11 taps in each of its 4 feedforward filters. The performance gain becomes much larger at higher $\frac{E_b}{N_0}$ where effects of ISI becoems more pronounced.

Delay (μsec)	0.0	0.2	0.5	1.6	2.3	5.0
Strength (dB)	-3.0	0.0	-2.0	-6.0	-8.0	-10.0

Table 1: EDGE TU Channel Model

5 Conclusions

We showed how to combine a multi–input multi–output MMSE–DFE with space–time block codes to combat frequency–selective fading. For Alamouti's space–time block code with 2 transmit and 2 receive antennas, the MIMO DFE consists of 4 feedforward FIR filters that perform diversity combining and combat MIMO pre–cursor intersymbol interference and 4 feedback filters that combat MIMO post–cursor intersymbol interference. A raw BER of 10^{-4} is achieved at $\frac{E_b}{N_0} \approx 18$ dB over a TU EDGE channnel with a total (feedforward and feedback) of 44 taps and only linear processing complexity (no trellis search).

References

[1] A. Duel-Hallen. "Equalizers for Multiple Input Multiple Output Channels and PAM Systems with Cyclostationary Input Sequences". *IEEE Journal on Selected Areas in Communications*, pages 630–639, April 1992.

[2] A. Duel-Hallen. "A Family of Multiuser Decision-Feedback Detectors for Asynchronous Code-Division Multiple-Access Channels". *IEEE Transactions on Communications*, pages 421–434, Feb/Mar/April 1995.

[3] N. Al-Dhahir and A.H.Sayed. "The Finite-Length MIMO MMSE-DFE". *IEEE Transactions on Signal Processing*, October 2000.

[4] N. Al-Dhahir and J.M.Cioffi. "MMSE Decision Feedback Equalizers : Finite-Length Results". *IEEE Transactions on Information Theory*, 41(4):961–976, July 1995.

[5] N. Al-Dhahir and Ali H. Sayed. "A Computationally-Efficient FIR MMSE-DFE for Multi-User Communications". In *Asilomar Conference on Signals, Systems, and Computers*, October 1999.

[6] S. Alamouti. "A Simple Transmit Diversity Technique for Wireless Communications". *IEEE Journal on Selected Areas in Communications*, pages 1451–1458, October 1998.

[7] A. Furuskar, S. Mazur, F. Muller, and H. Olofsson. "EDGE : Enhanced Data Rates for GSM and TDMA/136 Evolution". *IEEE Personal Communications Magazine*, June 1999.

[8] V. Tarokh, H. Jafarkhani, and A.R. Calderbank. "Space-Time Block Codes from Orthogonal Designs". *IEEE Transactions on Information Theory*, pages 1456–1467, July 1999.

[9] V. Tarokh, H. Jafarkhani, and A.R. Calderbank. "Space-Time Block Coding for Wireless Communications : Performance Results". *IEEE Journal on Selected Areas in Communications*, pages 451–460, March 1999.

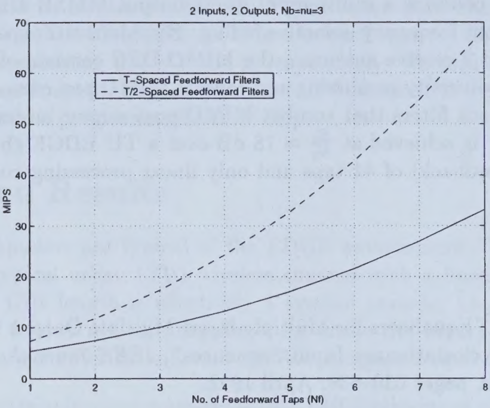


Figure 1: Instructions count for computing and applying MIMO MMSE-DFE coefficients versus N_f and l

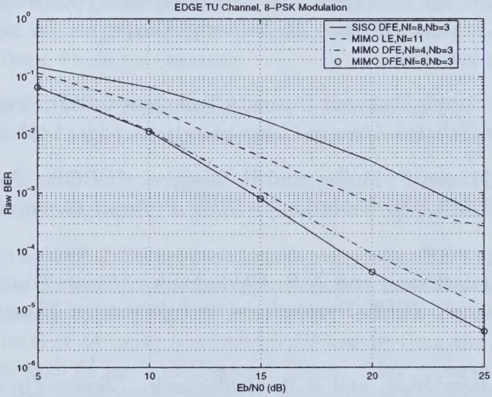


Figure 2: Raw BER of FIR MIMO MMSE-DFE, FIR MIMO MMSE-LE, SISO MMSE-DFE and Flat-Fading Scenarios

Codes for Unknown ISI Channels

Orhan Coskun and Keith M. Chugg
Communication Sciences Institute
Electrical Engineering–Systems Dept.
University of Southern California
Los Angeles, CA 90089-2565
coskun@usc.edu and chugg@usc.edu

Abstract

The traditional method of sending a training signal to identify a channel, followed by data, may be viewed as a simple code for the unknown channel. Results in blind sequence detection suggest that performance similar to this traditional approach can be obtained without training. However, for short packets and/or time-recursive algorithms, significant error floors exist due to the existence of sequences that are indistinguishable without knowledge of the channel. In this paper, we reconsider training signal design in light of recent results in blind sequence detection. We design *training codes* which combine modulation and training. In order to design these codes, we find an expression for the pairwise error probability of the joint maximum likelihood (JML) channel and sequence estimator. This expression motivates a pairwise distance for the JML receiver based on principal angles between the range spaces of data matrices. The general code design problem (generalized sphere packing) is formulated as the clique problem associated with an unweighted, undirected graph. We provide optimal and heuristic algorithms for this clique problem. For short packets, we demonstrate that significant improvements are possible by jointly considering the design of the training, modulation, and receiver processing.

1 Introduction and Summary

The traditional approach to communicating over an unknown intersymbol interference (ISI) channel is to send a training signal to identify the channel and then send information that is detected using this channel estimate. The training sequence is selected to have good auto-correlation properties and conveys no information in such a system. The other extreme is to attempt to detect the data "blindly." The blind approach eliminates overhead, but typically suffers from lost data during a period when the channel is acquired and/or misacquisition phenomena.

In this paper, we consider the overhead associated with channel identification to be part of the more general signal design problem. We establish a criterion for the design of block codes for unknown ISI channels under the assumption of joint maximum likelihood channel estimation and sequence detection (*i.e.*, JML for brevity). This criterion includes a relatively simple distance measure and a sphere-packing optimization algorithm. The distance measure results from determining the pairwise error probability for the JML detector and establishing a simple proxy distance that well approximates this probability. The sphere-packing algorithm is based on formulating the code design as a clique problem on a graph. This technique is demonstrated, conceptually, by designing short block codes that allow for reliable data transmission with low overhead.

The mathematical details of the clique algorithm suggested and the pairwise error probability expressions are contained in the two appendices.

1.1 Signal Model and Background

After appropriate front-end filtering and sampling, the received signal $\{z_i\}_{i=1}^k$ can be modeled as

$$\begin{aligned} \mathbf{z}_k &= A_k\,\mathbf{f} + \mathbf{w}_k & (1) \\ \mathbf{f} &= \begin{bmatrix} f_L \; f_{L-1} \; \cdots \; f_1 \end{bmatrix}^T & (2) \\ A_k &= \begin{bmatrix} a_{k-L+1} & \cdots & a_k \\ \vdots & & \vdots \\ a_{-L+2} & \cdots & a_1 \end{bmatrix} & (3) \end{aligned}$$

where $(\cdot)^T$ denotes transposition and $\mathbf{y}_k = \begin{bmatrix} y_k & y_{k-1} & \cdots & y_1 \end{bmatrix}^T$ is the notation for a signal vector to time k (*i.e.*, $\mathbf{z}_k$ and $\mathbf{w}_k$). The additive white Gaussian noise $\mathbf{w}_k$ is zero mean with covariance matrix $\sigma_w^2\mathbf{I}_k$. In the examples presented in this paper, a_k is drawn from $\{-1,+1\}$ which models binary phase shift keying (BPSK), but the development is applicable to arbitrary M constellations. Except for the length of the impulse response,[1] the continuous time channel is assumed to be unknown in this work, so that the model of (1) arises by some form of filtering and sampling of a continuous time observation.

In a trained system, the channel estimate is typically taken to be the least-squares estimate based on the model in (1). The associated estimation error variance is minimized by selecting a digital sequence such that the vector of singular values of A_k is as closely aligned to the all ones vector as possible. When the channel is unknown, the Viterbi algorithm is not applicable [1] and exhaustive search of possible sequences is required to implement the JML receiver which attempts to minimize

$$\Lambda_k(\tilde{A}_k) = \| \, (I - \tilde{A}_k\tilde{A}_k^I)\mathbf{z}_k \, \|^2 = \| \, (I - \tilde{P}_k)\mathbf{z}_k \, \|^2 \qquad (4)$$

over all possible allowable sequences $\tilde{A}_k$. The matrix $\tilde{A}_k^I$ is the pseudo-inverse of $\tilde{A}_k$ which is $(\tilde{A}_k^T\tilde{A}_k)^{-1}\tilde{A}_k^T$ for full-rank $\tilde{A}_k$ and $\tilde{P}_k = \tilde{A}_k\,\tilde{A}_k^I$ is the matrix that projects onto the range of $\tilde{A}_k$.

Note that, even with training, the JML receiver should perform an exhaustive search of all possible sequences – the initial training sequence simply eliminates a subset of sequences from the search. With sufficient training, however, the channel is well estimated after the training and the performance difference between the exhaustive search and a Viterbi or PSP-based approach should be small.

2 Training Code Design Using Separability and Distance Conditions

In this section, we describe a method to design codes for JML receivers (or good approximations thereof) such that the overhead is minimized for a given performance level. We consider the case with AWGN. Note that, even without noise, there exist sequences that cannot be distinguished [2, 3], hence the noise-free case is non-trivial. Given a set of pairwise distances, the design procedure is, however, the same for both cases. This design problem is one of sphere packing. Thus, we first describe our approach to the general code design problem, then we establish the appropriate distance measure and demonstrate the approach in the noisy case.

---

[1] The approaches discussed throughout are robust to overestimation of L, but sensitive to underestimation of L.

2.1 Code Design Methodology

We propose to design length N training codes by starting with all possible 2^N sequences and discarding sequences until the largest set of sequences meeting some pairwise condition is obtained. We then select the largest integer D such that 2^D is less than or equal to the size of this largest set, yielding a (N, D) code. In the noise-free case, this condition is separability, while with noise, it is some minimum distance requirement. Toward this goal, the condition used to discard sequences should (i) account for the digital nature of allowable sequences, (ii) be a pairwise property, (iii) be symmetric, (iv) be relatively simple to check, and (v) be independent of the actual channel coefficients $\mathbf{f}$.

Given some condition with the above properties, we can formalize the code design problem as finding the largest complete subgraph (or clique) on an undirected, unweighted graph – *i.e.*, the so-called *clique problem* [4]. Specifically, consider a graph representation of the code design problem with each of the 2^N potential codewords defining a vertex. An edge exists between two vertices if the associated sequences satisfy the condition. Due to the above properties, the resulting graph $G = (V, E)$ (*i.e.*, vertices V and edges E) is undirected and unweighted. The design problem is to find the largest subset $\mathcal{C}$ of the $|V| = 2^N$ vertices for which the pairwise condition is satisfied for each pair in $\mathcal{C}$.

The general class of clique problems is known to be NP-Hard [4]. However, in Appendix A we present an optimal algorithm which enumerates all cliques in an efficient manner when the maximum number of edges for any vertex is much smaller than $|V|$. We also describe a (suboptimal) greedy heuristic to approximate the solution to the clique problem for cases when the optimal algorithm is prohibitively complex.

2.2 A Pairwise Distance for JML Channel and Sequence Estimation

We introduce a distance criterion for JML that accounts for AWGN. The distance is related with the conditional second order statistics of the random variable r

$$r = \Lambda(A_k^{(j)}) - \Lambda(A_k^{(i)}) = \mathbf{z}_k^T \; (P_k^{(i)} - P_k^{(j)}) \; \mathbf{z}_k \tag{5}$$

In a pairwise JML decision between $A_k^{(i)}$ and $A_k^{(j)}$, the former is selected if and only if $r \geq 0$. In Appendix B, it is shown that, conditioned on $A_k^{(i)}$ being sent and a particular $\mathbf{f}$, r can be represented as a difference of two sums, each sum consisting of L mutually independent *non-central chi-square* random variables. It follows that the conditional probability density function (pdf) of r is the convolution of $2L$ densities of the χ^2 form. Therefore, for a given channel $\mathbf{f}$ and $A_k^{(i)}$, the pairwise error probability with $A_k^{(j)}$ can be found by numerical integration of this pdf function. A reasonable distance between $A_k^{(i)}$ and $A_k^{(j)}$ could be defined based on such pairwise error probabilities. However, this is numerically intensive and still should be made independent of the channel by averaging over all channels or considering the worst case.

We define a distance that is relatively simple to compute and its minimum value over unit norm channels is easily found. In Appendix B the conditional mean and the variance of r are shown to be

$$m_r(A_k^{(j)}|A_k^{(i)};\mathbf{f}) \triangleq \mathbb{E}\left\{r|A_k^{(i)};\mathbf{f}\right\} = \sum_{l=1}^{L} c_l^2 \sin^2\theta_l \qquad (6)$$

$$\sigma_r^2(A_k^{(j)}|A_k^{(i)};\mathbf{f}) \triangleq \text{var}\left[r|A_k^{(i)};\mathbf{f}\right] = 4\sum_{l=1}^{L} \sin^2\theta_l\sigma_w^4 + 4\sigma_w^2 m_r(A_k^{(j)}|A_k^{(i)};\mathbf{f}) \qquad (7)$$

where $\{\theta_i\}_{i=1}^{L}$ is the set of *principal angles* [5] between the range spaces of $A_k^{(i)}$ and $A_k^{(j)}$ and c_l is the coefficient vector $A_k^{(i)}\mathbf{f}$ represented with respect to a particular basis (see Appendix B).

Our distance is based on the value of $m_r(\cdot)/\sigma_r(\cdot)$. To motivate this choice, consider that if r were Gaussian the pairwise probability of error would be $\mathrm{Q}\left(m_r(\cdot)/\sigma_r(\cdot)\right)$. Thus, intuitively, a large value of $m_r(\cdot)/\sigma_r(\cdot)$ should reduce the conditional pairwise error probability. To illustrate this further, the conditional pairwise error rate (PER) was simulated for a pair of separable sequences as the channel traced the unit hemisphere. For all channels of length 3, two sequences of length 8 were sent 4×10^5 times through an AWGN channel with $\sigma_w^2 = 0.1$. Fig. 1 shows the simulation results of the PER versus $m_r(\cdot)/\sigma_r(\cdot)$. Also shown is the function $Q(m_r(\cdot)/\sigma_r(\cdot))$ which was empirically found to be a good approximate upper bound for the conditional PER. We performed similar experiments for 20 different sequence pairs of length 8 and 11 at channel lengths 3 and 4 respectively. In all cases similar results were obtained; $m_r(\cdot)/\sigma_r(\cdot)$ characterized the PER and the Q-function provided an good approximate upper bound. Although we have not shown this to be a valid upper bound, some intuitive justification is provided by the fact that the pdf of the decision statistic r is asymmetric around its mean with the heavier tail in the direction away from the decision boundary.

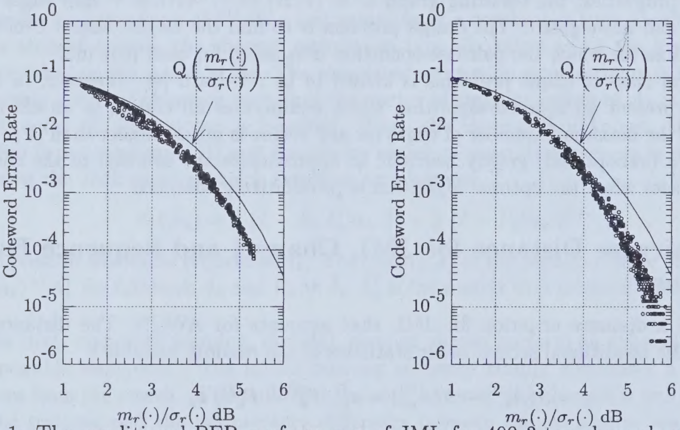


Figure 1: The conditional PER performance of JML for 400 3-tap channels selected on the unit hemisphere as a function of $m_r(\cdot)/\sigma_r(\cdot)$. Shown under the condition of either sequence transmitted. The empirical upper bound $\mathrm{Q}(m_r(\cdot)/\sigma_r(\cdot))$ is also shown.

The distance that we suggest is obtained by finding the smallest value of $m_r(\cdot)/\sigma_r(\cdot)$ considering all unit norm channels and either $A_k^{(i)}$ or $A_k^{(j)}$ as the transmitted sequence. It can be shown that $m_r(\cdot)/\sigma_r(\cdot)$ is an increasing function of $m_r(\cdot)$ so that the minimum value of $m_r(\cdot)/\sigma_r(\cdot)$ is achieved at the minimum value of $m_r(\cdot)$. While a simple lower bound on $m_r(\cdot)$ (see [6]) could be used to define a distance criteria, it is relatively simple to obtain the explicit minimum value of $m_r(\cdot)/\sigma_r(\cdot)$ for a unit norm channel. Let $\lambda_{\min}^{(i)}$ be the smallest eigenvalue of $A_k^{(i)}(P_k^{(i)} - P_k^{(j)})A_k^{(i)}$. This leads to the minimum value of the $m_r(\cdot)/\sigma_r(\cdot)$ and the minimum conditional distance

$$d(A_k^{(i)}|A_k^{(j)}) \triangleq \min_{\mathbf{f}:\|\mathbf{f}\|=1} \left[\frac{2m_r(A_k^{(j)}|A_k^{(i)};\mathbf{f})}{\sigma_r(A_k^{(j)}|A_k^{(i)};\mathbf{f})} \right] = \frac{\lambda_{\min}^{(i)}/\sigma_w^2}{\sqrt{\sum_{l=1}^{L} \sin^2\theta_l + \lambda_{\min}^{(i)}/\sigma_w^2}} \tag{8}$$

Finally, we define the distance as

$$d(A_k^{(i)}, A_k^{(j)}) = \min\left[d(A_k^{(i)}|A_k^{(j)}), d(A_k^{(j)}|A_k^{(i)}) \right] \tag{9}$$

so that the approximate upper bound for the pairwise error is $Q(d/2)$. The distance in (9) is an approximate minimax criterion because the minimum value of $m_r(\cdot)/\sigma_r(\cdot)$ approximately determines the maximum error probability. Note that the distance is a function of $\lambda_{\min}^{(i)}/\sigma_w^2$, which may be viewed as a measure of SNR for the quadratic detector.

2.2.1 Example of Noisy Design Methodology

We consider a simple example to illustrate the appropriateness of this distance. First, we design two sets of length $N = 8$ sequences, each of size three, for an $L = 3$ channel. These two sets are designed to be good or bad with respect to a known channel receiver or a JML receiver. Specifically, for a known channel $\mathbf{f}$, the pairwise error is monotonically decreasing in $\|(A_k^{(i)} - A_k^{(j)})\mathbf{f}\|$. Thus, analogous to the distance in (9) for the known channel case, we define a known-channel distance that is the minimum of $\|(A_k^{(i)} - A_k^{(j)})\mathbf{f}\|/(2\sigma_w^2)$ over all unit norm channels, which is simply minimum singular value of the matrix $A_k^{(i)} - A_k^{(j)}$ divided by $2\sigma_w^2$.

Set A was designed by setting the minimum value of the distance in (9) between any pair to 2.19 and the maximum known channel distance to 3.7. Similarly, Set B was found by setting the maximum distance in (9) between any pair to 1.6 and the minimum known channel distance to 4. All distances were computed at $\sigma_w^2 = 0.1$. The optimal clique algorithm was run on a graph of 128 vertices to determine the sets. The maximum clique size for each was found to be 3, so each set contains three sequences. Fig. 2 shows the average CER obtained by simulation of the JML receiver for the unknown channel case and the ML receiver for the known channel case. As expected, Set A is superior in the unknown channel case and inferior in the known channel case. This further validates the distance in (9) and distinguishes the approach from known channel signal design.

As another example, the distance criterion given in (9) was used to design a training-coded system for a packet length of $N = 16$. A codebook of size $2^7 = 128$ was found by running the greedy clique algorithm on a graph with 2^{15} vertices. Thus, a $(16, 7)$ training code was designed. Each sequence was sent 1000 times over all channels with the results summarized in Fig. 3. Unless specified, decoding was performed using the using JML via exhaustive search.

The performance of this combined coding and training system is compared with two traditional training systems. Trained system A uses the 9 bit overhead as an optimized

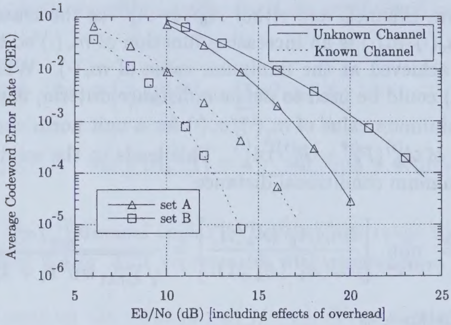


Figure 2: The CER for two sets of 3 sequences averaged over all channels. Set A was designed to have good unknown channel performance and set B was designed to have good known-channel performance.

training sequence at the beginning of the packet. This segregated design performs very poorly as compared to the combined design due to the two bits at the end of the packet. A second segregated design was considered which uses 7 bits of training at the beginning of the packet with the last two symbols of the packet fixed to terminate the channel state in a known manner. This trained system B performs much better than the first segregated design, but there is a 2.8 dB loss in SNR relative to the performance of the combined design at a CER value 10^{-3}. The performance of trained system B with segregated processing is also shown. Specifically, a standard (4 state) PSP receiver initialized by the 7 bit training sequence and terminated by the tail bits shows only slight degradation relative to the JML receiver in this case. Also shown in Fig. 3 is the performance of a known channel

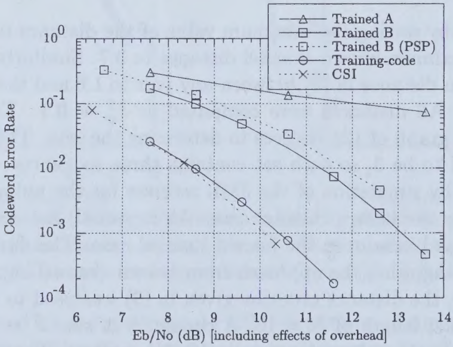


Figure 3: Comparison of segregated and combined training and modulation.

MLSD receiver averaged over all channels. This curve was generated by sending $N = 16$ bits with the first and last two bits specified to initialize and terminate the channel state, respectively. The performance of the system using the training code is similar to that of the perfect CSI system despite the SNR penalty due to an overhead increase (*i.e.*, 2.3 dB). This suggests that, in principle, one could obtain coding gain relative to an uncoded, known ISI channel while jointly estimating the channel. Realization of this gain most likely would require using larger block lengths which becomes prohibitively complex.

3 Concluding Remarks

Traditional training approaches may be viewed as simple codes for unknown channels that require relatively simple receiver processing at the expense of decreased rate or throughput. In this paper, we have established a framework for the design of training codes based on a pairwise distance that is appropriate for JML channel and sequence estimation or approximations thereof. For short data packet lengths the resulting training codes were demonstrated to provide significant improvements over traditional segmented approaches at the expense of increased receiver complexity. For longer packet lengths, the direct application of this framework becomes prohibitively complex.

Clearly, the major obstacle to realizing the potential gains for training codes using larger block lengths is the exponential complexity growth in both the code design and the decoding problems. An interesting area for future research is the formulation of the design in a recursive manner. Also, incorporation of additional levels of error correction coding using recent adaptive iterative detection techniques [7] is another promising research direction.

Appendices

A An Algorithm for Code Design (Clique Problem)

The general class of clique problems is known to be NP-Hard [4]. However, we present an optimal algorithm which enumerates all cliques in an efficient manner when the maximum number of edges for any vertex is much smaller than $|V|$. We describe this algorithm via the example in Fig. 4. The graph shown in Fig. 4 has $|V| = 16$ with a maximum number of edges for a given node of 7. A maximum clique is shown with a solid line and all other edges are shown dashed. The algorithm proceeds as follows:

1. Label the vertices of the graph from 0 to $|V| - 1$.

2. Let each vertex define the root of a tree. Construct the tree associated with vertex labeled v as follows. The children of the root v are all vertices with labels $\tilde{v}$ such that there is an edge between v and $\tilde{v}$, and $v < \tilde{v}$. For example, in Fig. 4, the tree with root label 3 has no child corresponding to label 2 because $3 > 2$. The edge between 2 and 3 has been represented in the tree rooted at label 2.

3. To grow each tree from the root level zero to level one, consider the child of v with smallest label, say $\tilde{v}$. The children of this node are determined by the intersection of all other children of root v with larger label than $\tilde{v}$ and the children of the tree rooted at $\tilde{v}$. For example, in Fig. 4, the children of 5 under the tree rooted at 2 are determined by taking the intersection of $\{6, 7, 8, 9, 10\}$ and the children of the root labeled 5, namely $\{6, 8, 9, 10\}$; yielding $\{6, 8, 9, 10\}$. The children of 6 under the root labeled 2 are determined by taking the intersection of $\{7, 8, 9, 10\}$ and the children of the root 6, namely $\{8, 10, 11, 12\}$; yielding $\{8, 10\}$.

4. At level $d > 0$, the tree is extended by the same process, except the set used for intersection can be simplified. Specifically, for any of the $|V|$ rooted trees, consider extending the vertex v at level d which has parent $P(v)$ at level $d-1$. The intersection of the set two sets provides the children of v. The first set is all vertices $\tilde{v}$ at level d in the same rooted tree with the same parent node $P(v)$ such that $v < \tilde{v}$. The second set is the children of the vertex label $\tilde{v}$ at level $d-1$ in the same rooted tree. For example, the children of vertex 6 at level 2 in the tree rooted by vertex 2 are obtained by intersecting $\{6, 8, 9, 10\}$ (set one), with the children of node 6 at level 1 in the same rooted tree, namely $\{8, 10\}$. Note that one could also use the children of 6 at level one of the tree rooted by vertex 6 for the second set, but this is less efficient.

5. This process is repeated on each of the rooted trees until it terminates. The largest clique is determined by the deepest level obtained in one of the trees. Notice that all cliques are enumerated in the collection of trees. In the example, two cliques of size 5 exist: $\{2, 5, 6, 8, 10\}$ from the root 2 and $\{4, 7, 10, 11, 12\}$ from the root 4.

The correctness of this algorithm can be formally proven. Also, the running time of this algorithm can be shown to be upper bounded as exponential in N_E, the maximum number of edges associated with any vertex and polynomial in $|V|$. By comparison, exhaustive search to check if a clique of size n exists has running time $\frac{N_E!}{(N_E-n)!n!}$. However, even though the complexity relative to exhaustive search is small, when N_E and/or $|V|$ becomes large, this algorithm becomes prohibitively complex.

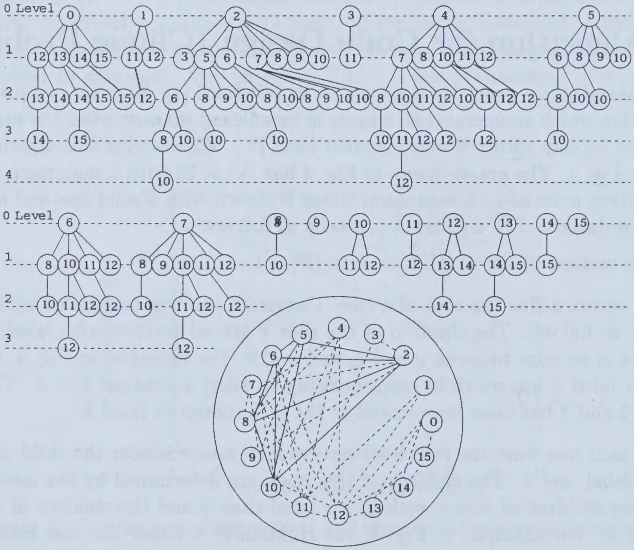


Figure 4: Example of the execution of the clique algorithm. One of the maximum cliques of the graph (in circle) is emphasized by non-dashed edges.

A.1 Greedy Heuristic for Code Design

We propose the following suboptimal algorithm when the above algorithm becomes prohibitively complex. Construct the set of $|V|$ trees as described above. At each level, grow only the tree with the largest number of children. For example, in Fig. 4, the tree rooted by 2 would be selected to be extended to level 1 because it has 7 children while no other root has more than 5 children. Only the vertex labeled 5 under the tree rooted at 2 would be extended to level two because it has 4 children while no other node at level two under the tree rooted at 2 has more than two children. When considering the next level there is a tie condition because both vertices 6 and 8 at level 2 have two children. In our heuristic an arbitrary tie breaker is used. If this breaks the tie in favor of 6, then the optimal solution will be found by this heuristic. However, if the tie-breaker selects 8 over 6 at level 2, then two cliques of size 4 are found (suboptimal) – *i.e.*, $\{2,5,8,9\}$ and $\{2,5,8,10\}$.

B PDF of the Quadratic Detector Signal in AWGN Environment

In this appendix we derive the probability density function (pdf) of the quadratic detector signal which is defined in (5). For compactness, we drop the time subscript k and do not explicitly denote dependence on the two data sequences and the actual channel for most quantities discussed.

Let $\{\mathbf{u}_1, ..., \mathbf{u}_L\}$ and $\{\mathbf{v}_1, ...\mathbf{v}_L\}$ be the principal vectors with angles $0 < \theta_1 \le \theta_2 < \cdots \le \theta_L$ between the column spaces of two separable data matrices, $A^{(i)}$ and $A^{(j)}$. A set of orthonormal eigenvectors of $P^{(i)} - P^{(j)}$ associated with nonzero eigenvalues can be written as

$$\mathbf{e}_l^{(+)} = \frac{1}{\sqrt{2}}\left(\frac{\mathbf{u}_l + \mathbf{v}_l}{||\mathbf{u}_l + \mathbf{v}_l||} - \frac{\mathbf{u}_l - \mathbf{v}_l}{||\mathbf{u}_l - \mathbf{v}_l||}\right) \tag{10}$$

$$\mathbf{e}_l^{(-)} = \frac{1}{\sqrt{2}}\left(\frac{\mathbf{u}_l + \mathbf{v}_l}{||\mathbf{u}_l + \mathbf{v}_l||} + \frac{\mathbf{u}_l - \mathbf{v}_l}{||\mathbf{u}_l - \mathbf{v}_l||}\right) \tag{11}$$

where $l = \{1, \cdots, L\}$. To show this, note that the denominators of (10) and (11) can be written in terms of principal angles,[2]

$$||\mathbf{u}_l \pm \mathbf{v}_l|| = \sqrt{2 \pm 2\cos\theta_l} \tag{12}$$

The product of the given vectors $\mathbf{e}_l^{(+)}$ and $\mathbf{e}_l^{(-)}$ by $P^{(i)}$-$P^{(j)}$ can be shown to be

$$(P^{(i)} - P^{(j)})\mathbf{e}_l^{(\pm)} = \pm\sin\theta_l \mathbf{e}_l^{(\pm)} \tag{13}$$

Thus, $\mathbf{e}_l^{(+)}$ and $\mathbf{e}_l^{(-)}$ are the eigenvectors of $P^{(i)} - P^{(j)}$. It follows that, if λ_l is an eigenvalue of $P^{(i)} - P^{(j)}$ corresponding to the eigenvector $\mathbf{e}_l^{(+)}$, then -λ_l is also an eigenvalue of $P^{(i)} - P^{(j)}$ corresponding to the eigenvector $\mathbf{e}_l^{(-)}$.

---

[2] For conciseness, we use ± and ∓ to express two equations in one using the standard convention – *e.g.*, $x^{(\pm)} = y \mp z$ is short for $x^{(+)} = y - m$ and $x^{(-)} = y + m$.

The principal vectors $\{\mathbf{u}_l\}_{l=1}^L$ and $\{\mathbf{v}_l\}_{l=1}^L$ can be expressed in terms of eigenvectors of $P^{(i)} - P^{(j)}$ as

$$\begin{aligned}
\mathbf{u}_l &= \frac{1}{2}\left(\sqrt{1+\cos\theta_l}+\sqrt{1-\cos\theta_l}\right)\mathbf{e}_l^{(+)} + \frac{1}{2}\left(\sqrt{1+\cos\theta_l}-\sqrt{1-\cos\theta_l}\right)\mathbf{e}_l^{(-)} && (14)\\
\mathbf{v}_l &= \frac{1}{2}\left(\sqrt{1+\cos\theta_l}-\sqrt{1-\cos\theta_l}\right)\mathbf{e}_l^{(+)} + \frac{1}{2}\left(\sqrt{1+\cos\theta_l}+\sqrt{1-\cos\theta_l}\right)\mathbf{e}_l^{(-)} && (15)
\end{aligned}$$

which follows by solving (10) and (11) for $\mathbf{u}_l$ and $\mathbf{v}_l$.

Theorem B.1 If the transmitted sequence is $A_k^{(i)}$, the channel is $\mathbf{f}$, and the AWGN variance is σ_w^2, then r can be expressed in terms of $2L$ *independent non-central chi-square distributed* random variables $\{r_l^{(+)}, r_l^{(-)}\}_{l=1}^L$

$$r = \sum_{l=1}^{L}(r_l^{(+)} - r_l^{(-)}) \qquad (16)$$

with mean and variances

$$\begin{aligned}
m_{r_l^{(\pm)}} &= \frac{1}{2}c_l^2(1\pm\sin\theta_l)\sin\theta_l + \sigma_w^2\sin\theta_l && (17)\\
\sigma^2_{r_l^{(\pm)}} &= 2\sigma_w^4\sin^2\theta_l + (\sigma_w\sin\theta_l c_l)^2(1\pm\sin\theta_l) && (18)
\end{aligned}$$

where $\{c_l\}_{l=1}^L$ are the coordinates of the vector $A_k^{(i)}\mathbf{f}$ with respect to the basis $\{\mathbf{u}_l\}_{l=1}^L$.

Proof. Using the spectral decomposition of $P^{(i)} - P^{(j)} = \sum_{l=1}^L \sin\theta_l(\mathbf{e}_l^{(+)}[\mathbf{e}_l^{(+)}]^T - \mathbf{e}_l^{(-)}[\mathbf{e}_l^{(-)}]^T)$, (5) can be written as

$$r = \sum_{l=1}^{L}([y_l^{(+)}]^2 - [y_l^{(-)}]^2) \qquad (19)$$

Since $\mathbf{z}_k$ is Gaussian and $\{\mathbf{e}_l^{(+)}, \mathbf{e}_l^{(-)}\}_{l=1}^L$ is an orthonormal set, $\{y_l^{(\pm)}\}$ are independent Gaussian random variables with means $\frac{1}{2}\sqrt{\sin\theta_l(1\pm\sin\theta_l)c_l^2}$ and variance $\sigma_w^2\sin\theta_l$. Therefore $\{r_l^{(\pm)} = [y_l^{(\pm)}]^2\}$ are *chi-square distributed* with means and variances given by (17) and (18), respectively.

Note that the conditional pdf of r can be expressed as the following $2L$-fold convolution

$$f_r(r|A^{(i)};\mathbf{f}) = f_{r_1^{(+)}}(r|A^{(i)};\mathbf{f}) * f_{r_1^{(-)}}(-r|A^{(i)};\mathbf{f}) \cdots * f_{r_L^{(+)}}(r|A^{(i)};\mathbf{f}) * f_{r_L^{(-)}}(-r|A^{(i)};\mathbf{f}) \qquad (20)$$

with the $2L$ densities defined by

$$f_{r_l^{(\pm)}}(r|A^{(i)};\mathbf{f}) = \frac{\exp\left[\frac{-r+c_l^2(1\pm\sin\theta_l)\sin\theta_l}{2\sigma_w^2\sin\theta_l}\right]}{\sqrt{2\pi r\sigma_w^2\,\sin\theta_l}}\cosh\left(\frac{\sqrt{rc_l^2(1\pm\sin\theta_l)\sin\theta_l}}{\sigma_w^2\sin\theta_l}\right) \qquad r \geq 0 \qquad (21)$$

This follows directly from the fact that $\{r_l^{(+)}, r_l^{(-)}\}$ is a set of independent chi-square random variables and the density of $-r_l^{(-)}$ is $f_{r_1^{(-)}}(-r_l|A^{(i)};\mathbf{f})$. Furthermore, the conditional mean and the variance of r are as stated in (6) and (7), respectively. Finally, it is straightforward to show that $m_r(\cdot)/\sigma_r(\cdot)$ is an monotonically increasing function of $m_r(\cdot)$.

References

[1] K. M. Chugg, "The condition for applicability of the Viterbi algorithm with implications for fading channel MLSD," *IEEE Trans. Commun.*, vol. 46, pp. 1112–1116, September 1998.

[2] F. Gustafsson and B.Wahlberg, "Blind equalization by direct examination of the input sequences," *IEEE Trans. Commun.*, vol. 43, pp. 2213–2222, July 1995.

[3] K. M. Chugg, "Blind acquisition characteristics of PSP-based sequence detectors," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 1518–1529, October 1998.

[4] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms.* McGraw-Hill, 1997.

[5] G. H. Golub and C. F. V. Loan, *Matrix Computation.* John Hopkins, 1996.

[6] O. Coskun and K. M. Chugg, "Combined coding and training for unknown isi channels." submitted to *IEEE Trans. Commun.*, June 2000.

[7] A. Anastasopoulos and K. M. Chugg, "Adaptive soft-input soft-output algorithms for iterative detection with parametric uncertainty," *IEEE Trans. Commun.*, 2000. (to appear).

Scattering of Ultra-Wideband Signals [†]

Toby Berger
School of Electrical and Computer Engineering
Cornell University
Ithaca, NY 14853
berger@ee.cornell.edu

Jay Cheng

School of Electrical and Computer Engineering
Cornell University
Ithaca, NY 14853
jcheng@ee.cornell.edu

Abstract

Given a communication channel of bandwidth B impaired both by fading and by additive gaussian noise, suppose one distributes a fixed amount P of average transmitter power effectively uniformly over the channel passband. Then Gallager and Medard [1] and Teletar and Tse [2] recently have shown that, provided certain additional assumptions concerning delay spread, coherence time, and the number of resolvable paths are met, the rate at which information can be transferred across the channel becomes vanishingly small in the limit as $B \to \infty$. By analyzing models of electromagnetic scattering from reflective surfaces in the physical channel, we deduce that not all the assumptions upon which this conclusion rests need necessarily be fulfilled in the ultra-wideband limit. Accordingly, spread spectrum signaling may not fail in this limit after all.

1 Ultra-Wideband RF Channels

In previous studies the carrier frequency f_c has been kept fixed during exploration of the consequences of increasing the bandwidth B. This may be a satisfactory mode of analysis for practical situations in which B is circa 10 MHz and f_c is in the GHz range; however, it clearly is not viable once B exceeds f_c and certainly is untenable in the limit as $B \to \infty$. The ultra-wideband limit is a theoretical abstraction. For instance, it ignores the intense attenuation at high frequencies in the presence of precipitation. Nonetheless, it is of theoretical interest and may be of practical interest in certain indoor situations.

Here is our ultra-wideband model. For $n = 1, 2, \ldots$ let $\mathcal{C}_n$ be an RF channel of bandwidth B_n and center frequency f_n corrupted both by fading and by additive gaussian noise. Each $\mathcal{C}_n$ is traditionally narrowband in that $B_n \ll f_n$, but the $\mathcal{C}_n$ become ultra-wideband as n increases in the sense that $B_n \to \infty$ and, of necessity, $f_n \to \infty$, as well.

[†] This research is supported in part by NSF Grant No. CCR-9980616

2 Delay Spread and Coherence Time

Let $\tau_l(t)$ denote the delay at time t of l^{th} propagation path from the transmitter to the receiver. Electromagnetic energy following this path may experience one or more reflections off scatterers in channel environment. The *delay spread* is given by the prescription

$$T_d = \sup_{l,m,t}(\tau_l(t) - \tau_m(t)).$$

Note that T_d does not depend on the carrier frequency f_c. It is a function solely of the configuration of the scatterers and the placement of the transmitting and receiving antennas.

The *coherence time* T_c is the maximum time separation over which the differential phase delay over any given path is only a negligible fraction of a cycle; specifically,

$$T_c = \max\{T : \sup_{\substack{l,s,t: \\ |s-t| \le T}} f_c[\tau_l(t) - \tau_l(s)] \ll 1\}.$$

If there is any motion whatsover of the transmitter, the recevier, and/or the scatterers, then it is obvious that $T_c \to 0$ as $f_c \to \infty$. Hence, whereas it is reasonable to assume in today's practical situations that $T_d \ll T_c$ [2], in our ultra-wideband limit the opposite inequality necessarily prevails. This opens the door to reconsideration of the nature of effective broadband signaling strategies in the said ultra-wideband limit.

3 One Line Scatterer

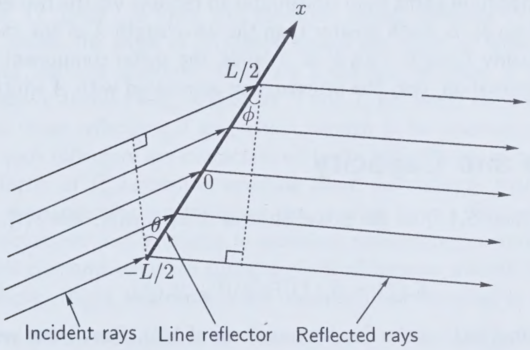


Figure 1: Incident and reflected rays for a line reflector of length L

Temporarily suppose that there is only one scatterer in the channel environment, an ideal line reflector R of length L as shown in Fig. 1. We consider that R is located in the far field of the transmitter. Hence, the transmitted waveform arrives at R as a plane wavefront incident at an angle we denote by θ. An exact electromagnetic scattering analysis would account for the singularities at the ends of the reflector and other considerations of antenna theory and field theory. We shall, however, settle for a simple ray model. The only reflected rays of interest to us are those that reach the receiver, which also is assumed to be in the far field of R. Consider the plane containing R and the line of sight S from the center of R to the focal point of the receiving antenna, and let ϕ denote complement of the angle between these two lines. Then the rays in question, labeled the reflected rays in Fig. 1, are the ones that lie in said plane, intersect R, and are parallel to S.

Denote the transmitted signal by $s(t)$ and the received signal by $r(t)$. Under our assumptions

$$r(t) = \int_{-L/2}^{L/2} As(t - \frac{x(\sin\theta - \sin\phi)}{c} - \Delta)\mathrm{d}x + n(t), \tag{1}$$

where A is the path gain, c is the speed of light, Δ is the delay of the ray passing through $x = 0$, and $n(t)$ is additive Gaussian noise. Without loss of generality, we can put $A = 1$ and $\Delta = 0$. Let $a = (\sin\theta - \sin\phi)/c$ and make a change of variable so that we can express $r(t)$ as

$$\begin{aligned} r(t) &= \frac{1}{a}\int_{-aL/2}^{aL/2} As(t-y)\mathrm{d}y + n(t) \\ &= s(t) * h(t) + n(t), \end{aligned} \tag{2}$$

where

$$h(t) = \frac{1}{a}\mathrm{rect}(\frac{t}{aL}). \tag{3}$$

Note that the line reflector acts as a filter whose impulse response function is $h(t)$. Further note that this filter significantly attenuates all energy that is at frequencies $f \gg (c/L)|\sin\theta - \sin\phi|^{-1}$. We assume that all the transmitted energy is confined to a band of the form $|f - f_c| \leq B/2$, where B is the signal bandwidth, f_c is the signal carrier frequency, and $B \ll f_c$. Accordingly, the filtering action will severely attenuate the transmitted signal if $f_c \gg (c/L)|\sin\theta - \sin\phi|^{-1}$. Equivalently, severe attenuation prevails if the difference in the lengths of the propagation paths from transmitter to receiver via the two ends of the reflector, namely $L|\sin\theta - \sin\phi|$, is much greater than the wavelength λ of the carrier. Conversely, if the opposite inequality $L|\sin\theta - \sin\phi| \ll \lambda$ holds, the signal component at the receiver will not suffer this attentuation, only the attentuation associated with A which we have set to 1.

4 Spectra and Capacity

Let $S_s(f)$, $S_n(f)$, and $S_r(f)$ be the power spectra of $s(t)$, $n(t)$, and $r(t)$, respectively. Then we have

$$S_r(f) = S_s(f)|H(f)|^2 + S_n(f), \tag{4}$$

where $H(f) = L\mathrm{sinc}(faL)$ is the Fourier transform of $h(t)$. Recall that we are assuming that the transmitted signal is confined to a frequency band of bandwith B centered at a carrier frequency f_c. Further assume that its average power is required not to exceed P. The channel capacity C is given by the Kolmogorov-Pinsker *water-pouring* prescription [3], namely

$$C = \frac{1}{2}\int_{\{f:|f-f_c|\leq B/2\}} \log\,(1 + \frac{(\nu - \frac{S_n(f)}{|H(f)|^2})^+}{\frac{S_n(f)}{|H(f)|^2}})\mathrm{d}f, \tag{5}$$

where ν satisfies

$$P = \int_{\{f:|f-f_c|\leq B/2\}} (\nu - \frac{S_n(f)}{|H(f)|^2})^+\mathrm{d}f, \tag{6}$$

5 $C \overset{?}{=} 0$

Equations (5) and (6) tell us that in the important special case $S_n(f) = N_0/2$, we should pour the "water" into the reciprocal of a sinc-squared function. If we are in a situation in which $L|\sin(\theta) - \sin(\phi)| \gg \lambda$ (i.e., a situation in which $f_c aL \gg 1$), then this sinc-squared will be oscillating rapidly over the band. Moreover, the envelope of said sinc varies as f^{-2} which is nearly constant at the value f_c^{-2} over the band. We see below that as a consequence it is indeed true that $C \to 0$ as $B, f_c \to \infty$. Specifically, the power gets distributed nearly uniformly over the band and we obtain

$$\begin{aligned} C &\approx \frac{1}{2}\int_{\{f:|f-f_c|\leq B/2\}} \log{(1 + \frac{P}{N_0 B}|H(f)|^2)} \mathrm{d}f \\ &\leq \int_{f_c-\frac{B}{2}}^{f_c+\frac{B}{2}} \log{(1 + \frac{PL^2}{N_0 B(\pi f aL)^2})} \mathrm{d}f, \end{aligned} \tag{7}$$

where we have upperbounded $sin(\cdot)^2$ by 1. This scarcely changes if we replace f by f_c througout the range of integration. We then find that the logarithm in the integrand nearly equals $PL^2/(N_0 B\pi^2 f_c^2 a^2 L^2)$. Integrating cancels the B in the denominator and yields

$$C \leq \frac{PL^2}{N_0(\pi f_c aL)^2}. \tag{8}$$

Since P, L and N_0 are constants and we are currently assuming that $f_c aL \gg 1$, this approaches zero as B, and hence f_c, grow without bound.

But what if $f_c aL \ll 1$? Since L is fixed, this condition cannot continue to prevail as $f_c \to \infty$ unless $a = 0$, i.e., unless the angle of incidence equals the angle of reflection. However, in practice there is not just one scatterer; there are a plethora of them of different sizes and orientations. The above analysis tells us that, as B and f_c get large, the channel environment reduces in effect to those reflectors, if any, which happen to be oriented in such a way that $\theta \approx \phi$. Moreover, such reflectors are characterized by an $h(t)$ that does not attenatuate the signal power by a factor of f_c^2 across the signaling band. Accordingly, there is the possibility that there will be some specular scattering to the receiver from a few reflectors located at disparate ranges, elevations and azimuths in spherical coordinates centered at the receiver. We have all seen this phenomenon when sunlight glints off faraway smooth surfaces, especially near dawn and twilight. Might this prevent the capacity from decaying to zero?

The casual answer to the above question is, "No, not really." The road to this answer involves observing that as f_c grows it becomes less less likely that a randomly chosen scatterer will have a length and orientation that satisfies $f_c aL \ll 1$. Indeed, the probability falls in proportion to $1/f_c$. The same fact holds if one replaces the line scatterers we have been assuming with planar facets, a more realistic model in most environments. Hence, if one distributes the transmitted power nearly uniformly over the signaling band, it still seems that the information transfer rate will decay to zero as B and hence f_c grow without bound.

However, another consideration remains. Note that, as the number of effective scatterers diminishes with increasing f_c, the angles from which the received power is coming become few and far between. This opens up the possibility that the receiver can adapt its antenna pattern in order to steer narrow beams in these directions only. The consequence of that would be that the effective noise power spectral density N_0 also would drop in proportion to the increasing bandwidth. Likewise, in multiaccess situations, which consitute the application of greatest interest in practice, the interference level (from other in-cell users in the case of an asynchronous reverse link or from out-of-cell transmitters in the case of a synchronous direct link) also would drop in proportion to the increasing bandwidth. Such a drop in noise and interference power levels would compensate for the loss in signal power. The resulting preservation of the S/N and S/I ratios would prevent the mutual information from decaying in the wideband signaling case, at least until the point at which electronic circuit noise takes over from background radiation as the principal source of additive noise and/or interference. In this regard it has always been understood [2] that the capacity of the wideband fading channel would equal that of the associated ideal bandliimited AWGN channel with the same received power provided the channel's state is able to be estimated accurately. As B and f_c increase, there is a sense in which this channel estimation problem becomes simpler because of the associated decrease in the effective number of scatterers. On the other hand, we have just noted that there is also a sense in which the the channel estimation problem becomes harder because we need to understand the channel not only temporally and spectrally but also spatially. When all these considerations are taken into account, the possibility remains that broadband signals, such as those currently in use in CDMA schemes that expand bandwidth via coding and spreading, may remain efficacious in ultra-wideband communications. Whether they do or not depends principally, we believe, on whether most of the environment's scattering cross-section is attributable to a small number of large-scale reflectors or a large number of small-scale reflectors. It also depends on the distribution of scatterer orientations and on other geometical considerations in the channel, at the transmitter, and at the receiver.

When the transmitter and/or receiver are in motion, the duration for which the return from a given reflector would remain specular diminishes. This, together with the associated doppler spreading, would further complicate the demanding task of training narrow receiving antenna beams on those reflective elements principally responsible for the bistatic scattering cross-section. The same is true if the scatterers are permitted to undergo translational movement. Particularly problematic is the realistic case in which the scatterers also undergo yaw and/or pitch rotational movement. (Roll around the axis perpendicular to the plane of the scatterer is relatively innocuous.) Such rotations would further diminish, perhaps drastically, the time the receiver has to adapt its beam forming pattern. The severity of this problem is related to the three-dimensional shape of the scatterers, being severe when they are nearly planar and becoming negligible as they tend toward sphericity. On the positive side, as f_c grows the number of wavelengths spanned by a handheld device grows in proportion, allowing formation of narrow receive beams. Also, since B is growing as well, a tremendous number of bits can be sent successfully per unit time during those possibly relatively limited time intervals over which the channel's spectral and spatial state is satisfactorily understood by the receiver.

References

[1] R. G. Gallager and M. Medard, "Bandwidth scaling for fading channels," *Proc. of the Int. Symp. on Information Theory*, Ulm, Germany, 1997, p. 471.

[2] I. E. Teletar and D. N. C. Tse, "Capacity and mutual information of wideband multipath fading channels," *IEEE Trans. Inform. Theory*, vol. 46, pp. 1384-1400, July 2000.

[3] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, New York: Wiley, 1991.

On Modeling Multi-Antenna Multipath Channels

Akbar M. Sayeed*

Department of Electrical and Computer Engineering
University of Wisconsin–Madison
1415 Engineering Drive
Madison, WI 53706
akbar@engr.wisc.edu, http://dune.ece.wisc.edu

Abstract

Accurate and tractable channel modeling is critical to realizing the full potential of antenna arrays in wireless communications. In this paper we propose a framework for modeling multi-antenna multipath channels based on the notion of virtual spatial angles. The virtual angles are fixed a priori and are determined by the number of antennas at the transmitter and receiver and the spacing between the antennas. The model essentially corresponds to a coordinate transformation via fixed spatial basis functions (array response and steering vectors) at the transmitter and receiver. The resulting linear channel matrix representation in the virtual space encompasses all existing models and provides a natural link between the actual physical propagation environment and the channel statistics induced by it. For any given scattering environment, the model clearly reveals the two key parameters affecting channel capacity: the number of parallel channels and the level of diversity associated with each parallel channel. In addition to facilitating realistic estimates of channel capacity, the virtual space framework also suggests low-complexity channel approximations that may be leveraged in efficient transceiver design.

1 Introduction

The use of antenna arrays holds great promise for bandwidth-efficient communication over the harsh wireless channel. The advantages of antenna arrays are many-fold. They provide additional diversity to combat fading that can be exploited via appropriate space-time signaling and coding to enhance system capacity. They can be used for spatially selective signaling and reception to mitigate multiuser interference. Finally, antenna arrays also provide additional dimensions/degrees of freedom that can be exploited for efficient allocation of network resources, such as power and bandwidth, to accommodate disparate requirements of different users. All these attractive features have motivated a great deal of research into the use of antenna arrays in mobile wireless communications.

Maximal exploitation of antenna arrays in wireless communication necessitates accurate yet computationally tractable modeling of the multi-input multi-output (MIMO) channel coupling the transmitters and receivers. Existing models represent two extreme approaches. On the one hand are *statistical* models that are idealized abstractions of spatial propagation characteristics and assume independent fading between different

---

*This research was supported in part by the National Science Foundation under CAREER grant CCR-9875805 and grant ECS-9979408, and by the Wisconsin Alumni Research Foundation.

transmit-receive element pairs. These models have been heavily used in capacity calculations and development of space-time coding techniques (see, e.g., [1, 2, 3]). On the other hand are *physical* models, inspired by array processing techniques, that explicitly model signal copies arriving from different directions. While more accurate descriptions of the actual propagation environment, these models are nonlinear in spatial angles, thereby making it rather difficult to incorporate them in transceiver design and capacity computations. Finally, the two approaches to MIMO antenna array modeling exist in virtual isolation without explicit connections relating them. A connection is very much desirable so that conclusions and insights derived from the two approaches can be combined and leveraged into improved transceiver designs.

There have been some recent attempts at bridging the gap between the two modeling philosophies (see, e.g., [4, 5]). However, these approaches are heavily entrenched in the physical modeling paradigm and suffer from the drawbacks associated with nonlinear modeling. We propose an approach, inspired by modeling physical propagation characteristics in the context of spatial signal space dimensions, that keeps the essence of physical modeling without its complexity and imposes a natural structure on channel statistics. Specifically, we model the channel with respect to fixed spatial basis functions defined by fixed virtual angles. This yields a linear channel characterization that is significantly more tractable analytically and also facilitates tackling various issues with ease. For instance, it explicitly reveals the level of spatial diversity afforded by the channel, a factor that critically affects outage capacity. In the context of spatio-temporal signal processing, the model provides a natural approach for spatially selective signaling at the transmitter and low-complexity processing at the receiver. Finally, despite its simplicity, the model encompasses *all* existing approaches to modeling spatial multipath channels.

The next section introduces the three signal spaces for describing a multi-antenna multipath channel: the element space (statistical model), the physical space, and the proposed virtual space. Section 3 discusses various aspects of the virtual space model, including relation to the other two models and statistical channel characterization. Section 4 illustrates the utility of the proposed channel model via outage capacity calculations in a variety of scenarios. Section 5 provides some concluding remarks about extensions.

2 Element, Physical, and Virtual Spaces

Consider a transmitter array with n_T elements and a receiver array with n_R elements. In the absence of noise, the transmitted and received signals are related as

$$\boldsymbol{x} = \boldsymbol{H}\boldsymbol{s} \tag{1}$$

where $\boldsymbol{s}$ is the n_T dimensional transmitted signal, $\boldsymbol{x}$ is the n_R dimensional received signal and $\boldsymbol{H}$ denotes the channel matrix coupling the transmitter and receiver elements. This is the *element space* (ES) description of the spatial channel. Most capacity calculations use this model and assume $\boldsymbol{H}$ to consist of independent, identically distributed (iid) Gaussian random variables — an idealized, rich scattering environment.

Explicit modeling of physical propagation effects can be used to impose structure on $\boldsymbol{H}$ as illustrated in Figure 1. The most widely used *physical space* (PS) model is

$$\boldsymbol{H} = \sum_{l=1}^{L} \beta_l \boldsymbol{a}_R(\phi_{R,l})\boldsymbol{a}_T^H(\phi_{T,l}) = \boldsymbol{A}_R(\underline{\phi}_R)\boldsymbol{H}_P\boldsymbol{A}_T^H(\underline{\phi}_T) \tag{2}$$

which couples the transmitter and the receiver via L propagation paths with $\{\phi_{T,l}\}$ and $\{\phi_{R,l}\}$ as the spatial angles seen by the transmitter and receiver, respectively, and $\{\beta_l\}$ as the corresponding fading gains. In (2), $\boldsymbol{a}_R(\phi_R)$ represents the array response vector for a plane wave arriving from the direction ϕ_R and $\boldsymbol{a}_T(\phi_T)$ represents the transmitter steering vector for the direction ϕ_T. The angles $\{\phi_{R,l}\}$ lie in within S_R, the angular spread of

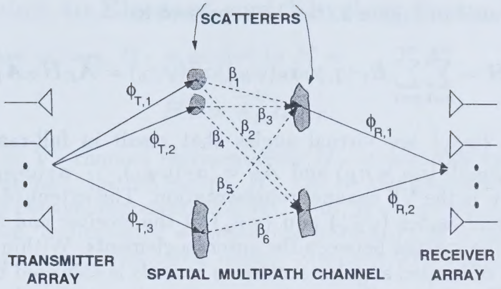

SCATTERERS
β_1
β_3
β_2
β_4
β_5
β_6
$\phi_{T,1}$
$\phi_{T,2}$
$\phi_{T,3}$
$\phi_{R,1}$
$\phi_{R,2}$
TRANSMITTER ARRAY
SPATIAL MULTIPATH CHANNEL
RECEIVER ARRAY

Figure 1: *A schematic illustrating channel modeling in the* **physical space**. *Each scattering path is associated with a fading gain (β_l) and a unique pair of transmit and receive angles ($\phi_{T,l}$, $\phi_{R,l}$) that connect physical scatterers arbitrarily distributed within the angular channel spreads.*

scatterers as seen by the receiver, whereas $\{\phi_{T,l}\}$ lie within S_T, the angular spread of scatterers as seen by the transmitter. In the matrix representation in (2), $\boldsymbol{A}_R(\underline{\phi}_R) = [\boldsymbol{a}_R(\phi_{R,1}), \cdots, \boldsymbol{a}_R(\phi_{R,L})]$ is an $n_R \times L$ matrix, $\boldsymbol{A}_T(\underline{\phi}_T) = [\boldsymbol{a}_T(\phi_{T,1}), \cdots, \boldsymbol{a}_T(\phi_{T,L})]$ is an $n_T \times L$ matrix, and $\boldsymbol{H}_P = \mathrm{diag}(\beta_1, \cdots, \beta_L)$ is an $L \times L$ diagonal matrix. The PS model (2) decomposes the channel matrix $\boldsymbol{H}$ in terms of $\boldsymbol{A}_R(\underline{\phi}_R)$, $\boldsymbol{A}_T(\underline{\phi}_T)$, and $\boldsymbol{H}_P$, all of which depend on actual physical parameters. The model is *linear* in the channel gains $\{\beta_l\}$ but *nonlinear* in the spatial angles $\{\phi_{R,l}, \phi_{T,l}\}$.

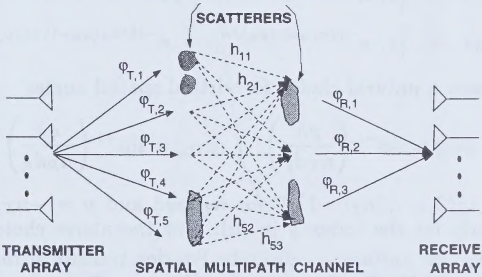

SCATTERERS
h_{11}
h_{21}
$\varphi_{T,1}$
$\varphi_{T,2}$
$\varphi_{T,3}$
$\varphi_{T,4}$
$\varphi_{T,5}$
h_{52}
h_{53}
$\varphi_{R,1}$
$\varphi_{R,2}$
$\varphi_{R,3}$
TRANSMITTER ARRAY
SPATIAL MULTIPATH CHANNEL
RECEIVER ARRAY

Figure 2: *A schematic illustrating channeling modeling in the* **virtual space**. *The virtual angles are fixed a priori and their spacing is determined by the antenna spacing and defines the spatial resolution — VS channel representation does not distinguish between scatterers that are within the spatial resolution; the scatterers corresponding to the physical angles $\phi_{T,1}$ and $\phi_{T,2}$, for example. The channel is characterized by the coefficients, $\{H_V(p,q) = h_{p,q}\}$, that couple the n_T virtual transmit angles, $\{\varphi_{T,p}\}$, with the n_R virtual receive angles, $\{\varphi_{R,q}\}$. In particular, for angles where there is no significant scattering at the transmitter and/or receiver, the corresponding coefficients are approximately zero (e.g., h_{21} and h_{52}).*

The finite dimensionality of the spatial signal space (due to the finite number of antenna elements) can be exploited to develop a *linear* channel model which uses spatial beams in *fixed, virtual* directions at the transmitter and the receiver. The *virtual space*

(VS) model, illustrated in Figure 2, can be expressed as

$$\boldsymbol{H} = \sum_{q=1}^{n_R} \sum_{p=1}^{n_T} H_V(q,p)\boldsymbol{a}_R(\varphi_{R,q})\boldsymbol{a}_T^H(\varphi_{T,p}) = \widetilde{\boldsymbol{A}}_R \boldsymbol{H}_V \widetilde{\boldsymbol{A}}_T^H \tag{3}$$

where $\{\varphi_{R,q}\}$ and $\{\varphi_{T,p}\}$ are virtual angles that result in full-rank matrices $\widetilde{\boldsymbol{A}}_R = [\boldsymbol{a}_R(\varphi_{R,1}), \cdots, \boldsymbol{a}_R(\varphi_{R,Q})]$ $(n_R \times n_R)$ and $\widetilde{\boldsymbol{A}}_T = [\boldsymbol{a}_T(\varphi_{T,1}), \cdots, \boldsymbol{a}_T(\varphi_{T,P})]$ $(n_T \times n_T)$. The $n_R \times n_T$ matrix $\boldsymbol{H}_V$ is the VS channel representation. The extent of the spatial horizon covered by the virtual angles $\{\varphi_{R,q}\}$ and $\{\varphi_{T,p}\}$ at the receiver and transmitter, respectively, depends on the spacing between the antenna elements. Within the covered spatial horizons, the effect of limited angular *scattering* spreads is captured by the non-zero values of the matrix $\boldsymbol{H}_V$. For smaller angular spreads, only certain sub-matrices of $\boldsymbol{H}_V$ are nonvanishing. In contrast to the PS model (2), the VS model is *linear* and is characterized by the matrix $\boldsymbol{H}_V$ due to the fixed choice of spatial angles. However, in contrast to $\boldsymbol{H}_P$, the matrix $\boldsymbol{H}_V$ is not diagonal in general. Finally, we note that the PS model has $3L$ parameters $\{\beta_l, \phi_{R,l}, \phi_{T,l}\}$, whereas the VS model has $n_T n_R$ parameters (entries of $\boldsymbol{H}_V$).

3 The Virtual Space Model

For simplicity of exposition, we develop the VS model in the context of uniform linear arrays (ULAs) at both the transmitter and the receiver with antenna spacings d_R and d_T, respectively. Analogous formulations for arbitrary geometries can also be developed. In the case of ULAs, the array steering and response vectors take the form

$$\boldsymbol{a}_T(\phi_T) = \left[1,\ e^{-j2\pi d_T \sin(\phi_T)/\lambda}, \cdots, e^{-j2\pi d_T (n_T-1)\sin(\phi_T)/\lambda}\right]^T \tag{4}$$

$$\boldsymbol{a}_R(\phi_R) = \left[1,\ e^{-j2\pi d_R \sin(\phi_R)/\lambda}, \cdots, e^{-j2\pi d_R (n_R-1)\sin(\phi_R)/\lambda}\right]^T \tag{5}$$

The following represent a natural choice for virtual spatial angles

$$\varphi_{T,p} = \sin^{-1}\left(\frac{p\lambda}{n_T d_T}\right), \quad \varphi_{R,q} = \sin^{-1}\left(\frac{q\lambda}{n_R d_R}\right) \tag{6}$$

where $p = -(n_T-1)/2, \cdots, (n_T-1)/2$ for n_T odd and $p = -n_T/2, \cdots, n_T/2 - 1$ for n_T even, and similarly for the index q in (6). For the above choice of angles, the VS matrices $\widetilde{\boldsymbol{A}}_R$ and $\widetilde{\boldsymbol{A}}_T$ are unitary — discrete Fourier transform matrices, in fact. We note that for $d_T = d_R = \lambda/2$, the virtual angles (6) cover the entire $([-\pi/2, \pi/2])$ spatial horizon. However, larger antenna spacings (covering smaller spatial horizons) may be used in spatial multiplexing applications.

To illustrate the effect of angular scattering spreads on $\boldsymbol{H}_V$, consider a single cluster of contiguous scatters with $S_R = [S_{R-}, S_{R+}] \subset [-\pi/2, \pi/2]$ and $S_T = [S_{T-}, S_{T+}] \subset [-\pi/2, \pi/2]$. In this case, $\boldsymbol{H}_V$ consists of a nonvanishing submatrix corresponding to S_R and S_T. In particular, $H_V(p,q)$ are nonzero for $q = Q_-, \cdots, Q_+$, $p = P_-, \cdots, P_+$ where $Q_- \approx \lfloor n_R \sin(S_{R-}) d_R/\lambda \rfloor$, $Q_+ \approx \lceil n_R \sin(S_{R+}) d_R/\lambda \rceil$, and similarly for P_- and P_+. For example, for $S_{R-} = -\pi/2$, $S_{R+} = \pi/2$ (maximum spread) and $d_T = d_R = \lambda/2$, $Q_- \approx -n_R/2$ and $Q_+ \approx n_R/2$ (all receive dimensions non-zero). The $\boldsymbol{H}_V$ corresponding to multiple clusters [4] is a superposition of nonvanishing submatrices corresponding to the scattering spreads associated with each cluster.

$\boldsymbol{H}_V$ provides a natural link between PS modeling via $\boldsymbol{H}_P$ and ES modeling via $\boldsymbol{H}$. We can establish this link by relating $\boldsymbol{H}_V$ to the PS matrix $\boldsymbol{H}_P$ and imposing an appropriate statistical model on $\boldsymbol{H}_V$. In particular, we demonstrate that by imposing a progressively more complex structure on $\boldsymbol{H}_V$, we can capture a fairly rich class of ES statistical models with a corresponding physical/propagation interpretation in the PS.

3.1 Relationship to Element and Physical Space Models

Since $\widetilde{\boldsymbol{A}}_R$ and $\widetilde{\boldsymbol{A}}_R$ are unitary, $\boldsymbol{H}_V$ is related to $\boldsymbol{H}$ as

$$\boldsymbol{H}_V = \widetilde{\boldsymbol{A}}_R^H \boldsymbol{H} \widetilde{\boldsymbol{A}}_T. \tag{7}$$

This implies that the VS channel representation $\boldsymbol{H}_V$ is *unitarily equivalent* to the ES representation $\boldsymbol{H}$. In fact, the VS matrix $\boldsymbol{H}_V$ can be thought of as a Fourier domain representation of ES channel representation $\boldsymbol{H}$. Thus, $\boldsymbol{H}_V$ captures all channel information and may be used, equivalently, wherever the ES matrix $\boldsymbol{H}$ is used — for example, for capacity calculations and transceiver design. Similarly, $\boldsymbol{H}_V$ is related to $\boldsymbol{H}_P$ as

$$\boldsymbol{H}_V = \widehat{\boldsymbol{A}}_R(\underline{\phi}_R)\boldsymbol{H}_P\widehat{\boldsymbol{A}}_T^H(\underline{\phi}_T) \tag{8}$$

where $\widehat{\boldsymbol{A}}_R(\underline{\phi}_R) = \widetilde{\boldsymbol{A}}_R^H \boldsymbol{A}_R(\underline{\phi}_R) = [\widehat{\boldsymbol{a}}_R(\phi_{R,1}), \cdots, \widehat{\boldsymbol{a}}_R(\phi_{R,L})]$ is a $n_R \times L$ matrix whose columns represent the projection of the PS array response vectors onto the VS array response vectors and $\widehat{\boldsymbol{A}}_T(\underline{\phi}_T) = \widetilde{\boldsymbol{A}}_T^H \boldsymbol{A}_T(\underline{\phi}_T) = [\widehat{\boldsymbol{a}}_T(\phi_{T,1}), \cdots, \widehat{\boldsymbol{a}}_T(\phi_{T,L})]$ is a $n_T \times L$ matrix whose columns represent the projection of the PS steering vectors onto the VS steering vectors. As illustrated in Figure 3, the vector $\widehat{\boldsymbol{a}}_R(\phi_{R,l})$ peaks at the few virtual angles in the neighborhood of the physical angle $\phi_{R,l}$ and a similar interpretation holds for $\widehat{\boldsymbol{a}}_T(\phi_{T,l})$.

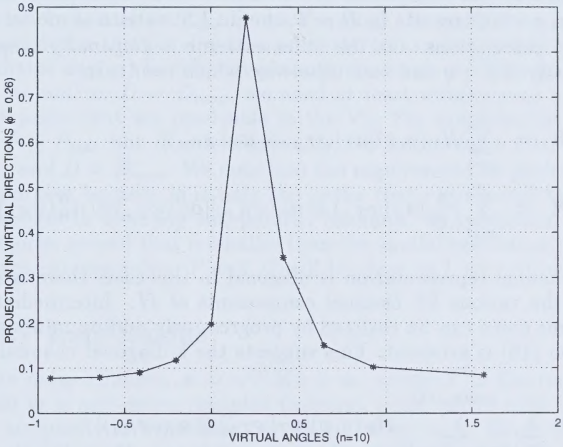


Figure 3: *Magnitude of the projection of a PS array response vector* $\boldsymbol{a}(\phi)$ *onto the VS array response vectors:* $|\widehat{\boldsymbol{a}}(\phi)| = |\widetilde{\boldsymbol{A}}^H \boldsymbol{a}(\phi)|$. *In this case,* $n = n_T = n_R = 10$. *Note that the VS projections peak near the actual physical angle* $\phi = 0.26$. *Though not illustrated here, the peaks get sharper with larger* n *(larger signal space dimensions).*

3.2 Statistical Channel Modeling in the Virtual Space

In this section, we discuss capturing the statistics of the ES channel matrix $\boldsymbol{H}$ via statistical modeling of $\boldsymbol{H}_V$. From (3) it follows that $\boldsymbol{h} = \text{vec}(\boldsymbol{H}) = [\widetilde{\boldsymbol{A}}_T \otimes \widetilde{\boldsymbol{A}}_R]\boldsymbol{h}_V$, where $\boldsymbol{h}_V = \text{vec}(\boldsymbol{H}_V)$, $\text{vec}(\boldsymbol{H})$ represents a vector obtained by stacking the columns of $\boldsymbol{H}$ and $\otimes$ represents the kronecker product [6]. Here we have used the identity $\text{vec}(\boldsymbol{ADB}) =$

$[\boldsymbol{B}^H \otimes \boldsymbol{A}]\text{vec}(\boldsymbol{D})$ [6]. Let $\boldsymbol{R} = \text{E}[\boldsymbol{h}\boldsymbol{h}^H]$ denote the correlation matrix $\boldsymbol{h}$ and let $\boldsymbol{R}_V$ denote the correlation matrix of $\boldsymbol{h}_V$. The two correlation matrices are related by $\boldsymbol{R} = [\widetilde{\boldsymbol{A}}_T \otimes \widetilde{\boldsymbol{A}}_R]\boldsymbol{R}_V[\widetilde{\boldsymbol{A}}_T^H \otimes \widetilde{\boldsymbol{A}}_R^H]$. $\boldsymbol{R}_V$ is always approximately diagonal[1] but may have some zero diagonal elements due to the sparseness of $\boldsymbol{H}_V$. We note that $\boldsymbol{R}$ and $\boldsymbol{R}_V$ are also unitarily equivalent.

We can obtain a more explicit characterization of $\boldsymbol{h}$ and $\boldsymbol{R}$ from (3)

$$\boldsymbol{h} = \sum_{q=1}^{n_R}\sum_{p=1}^{n_T} H_V(q,p)[\boldsymbol{a}_T(\varphi_{T,p}) \otimes \boldsymbol{a}_R(\varphi_{R,q})] \tag{9}$$

$$\boldsymbol{R} = \sum_{q=1}^{n_R}\sum_{p=1}^{n_T} \sigma_{q,p}^2[\boldsymbol{a}_T(\varphi_{T,p}) \otimes \boldsymbol{a}_R(\varphi_{R,q})][\boldsymbol{a}_T^H(\varphi_{T,p}) \otimes \boldsymbol{a}_R^H(\varphi_{R,q})] \tag{10}$$

where we have assumed that $\{H_V(q,p)\}$ are uncorrelated and $\sigma_{q,p}^2 = \text{E}[|H_V(q,p)|^2]$ is the power in each virtual channel coefficient. From (10) we note that the VS provides an eigen decomposition of $\boldsymbol{R}$, with $\{\boldsymbol{a}_T(\varphi_{T,p}) \otimes \boldsymbol{a}_R(\varphi_{R,q})\}$ serving as the eigenvectors and $\{\sigma_{q,p}^2\}$ the corresponding eigenvalues. (9) is the corresponding Karhunen-Loeve representation.

The relation (10) between $\boldsymbol{R}$ and $\{\sigma_{q,p}^2\}$ provides a systematic way of relating the statistics of the ES model to the actual physical scattering. Suppose that $n_T = n_R = n$ for simplicity of exposition. The $\{\sigma_{q,p}^2\}$ in (10) that are nonzero represent active scattering between the p^{th} transmit element and q^{th} receive virtual element. For simplicity assume that $\sigma_{q,p}^2$ is either 0 or 1. The extreme case of maximally rich scattering corresponds to $\sigma_{q,p}^2 = 1$ for all p, q which results in $\boldsymbol{R} = \boldsymbol{I}$, the iid ES statistical model usually assumed in most capacity calculations. On the other extreme is *diagonally coupled* scattering in which $\sigma_{q,p}^2 = 1$ only if $p = q$ and zero otherwise which results in

$$\boldsymbol{h} = \sum_{p=1}^{n} H_V(p,p)[\boldsymbol{a}_T(\varphi_{T,p}) \otimes \boldsymbol{a}_R(\varphi_{R,p})] \tag{11}$$

$$\boldsymbol{R} = \sum_{p=1}^{n} \sigma_{p,p}^2[\boldsymbol{a}_T(\varphi_{T,p}) \otimes \boldsymbol{a}_R(\varphi_{R,p})][\boldsymbol{a}_T^H(\varphi_{T,p}) \otimes \boldsymbol{a}_R^H(\varphi_{R,p})]. \tag{12}$$

While the VS channel representation is diagonal in this case, there is significant correlation between the various ES channel components of $\boldsymbol{H}$. Intermediate cases between these two extreme cases can be realized by progressively adding minor diagonals to (12) until the relation (10) is attained. This suggests the k-diagonal channel model[2]

$$\boldsymbol{h}_k = \sum_{p=1}^{n} \sum_{q=\max(1,p-k)}^{\min(n,p+k)} H_V(q,p)[\boldsymbol{a}_T(\varphi_{T,p}) \otimes \boldsymbol{a}_R(\varphi_{R,q})] \tag{13}$$

$$\boldsymbol{R}_k = \sum_{p=1}^{n} \sum_{q=\max(1,p-k)}^{\min(n,p+k)} \sigma_{q,p}^2[\boldsymbol{a}_T(\varphi_{T,p}) \otimes \boldsymbol{a}_R(\varphi_{R,q})][\boldsymbol{a}_T^H(\varphi_{T,p}) \otimes \boldsymbol{a}_R^H(\varphi_{R,q})] \tag{14}$$

where $0 \leq k \leq n-1$. As demonstrated in the next section, this approach for modeling progressively rich scattering in the VS provides a natural link between PS and ES modeling and is very useful in channel capacity calculations. In particular, (10) and (12) have nearly identical ergodic capacities (under appropriate power normalization) but starkly different outage capacities due to higher diversity in (10).

---

[1] Under the assumption of uncorrelated scattering as a function of spatial angle.

[2] Note that this also represents the orthogonal projection of $\boldsymbol{H}$ onto the space of k-diagonal $\boldsymbol{H}_V$'s since $\widetilde{\boldsymbol{A}}_R$ and $\widetilde{\boldsymbol{A}}_T$ are unitary.

4 Capacity Calculations in the Virtual Space

In this section, we discuss channel capacity calculations to illustrate the ease of computation and simple interpretation afforded by the VS model. We first briefly discuss the key parameters that control capacity and relate them to physical propagation characteristics. Our focus here is on outage capacity, a metric more appropriate in fading channels.

4.1 Parallel Channels and Diversity

A closer look at (8) reveals that, in essence, $\boldsymbol{H}_V$ provides an "image" of the physical scattering geometry. Recall that $\boldsymbol{H}_P$ is a diagonal matrix. If the physical angles $\{\phi_{R,l}\}$ and $\{\phi_{T,l}\}$ follow similar trends as a function of the path index (monotonically increasing or decreasing), then $\boldsymbol{H}_V$ would also follow a similar diagonal trend with a small spread due to the projection of the physical angles onto the virtual angles. On the other hand, if the coupling between the transmit and receive angles is richer (more cross-coupling between the different transmit and receive virtual angles), it will manifest itself as more nondiagonal terms in $\boldsymbol{H}_V$.

The "image" of the scattering environment provided by $\boldsymbol{H}_V$ is intimately related to two key parameters that affect channel capacity: the number of available *parallel channels*, P, that controls ergodic capacity, and the level of *diversity per parallel channel*, D, that controls the outage capacity. $P \leq P_{max} = \min(n_T, n_R)$, the rank of $\boldsymbol{H}$, and there have to be at least $L = \min(n_T, n_R)$ distinct paths (resolvable in the virtual space) to achieve P_{max}. However, to get maximum diversity, more paths are needed. The maximum diversity per parallel channel is $D_{max} = \max(n_T, n_R)$. In order to achieve D_{max}, each virtual angle associated with a parallel channel at the transmitter must be associated with D_{max} different virtual angles via different paths at the receiver. Thus, in order to achieve both $P = P_{max}$ as well as $D = D_{max}$, we need at least $\min(n_T, n_R) \times \max(n_T, n_R) = n_T \times n_R$ distinct paths that are resolvable in the VS. For example, the diagonal model in (12) achieves $P = P_{max}$ but $D = 1$ whereas the full iid model in (10) with all $\sigma^2_{p,q} = 1$ attains $P = P_{max}$ and $D = D_{max}$. We note that the requirement for paths to be resolvable in the VS is important because it is only the paths that correspond to distinct virtual angles that contribute to diversity and parallel channels. In particular, if the paths are confined to an angular spread that is smaller than the spatial resolution (spacing between virtual angles), the corresponding P and D will be close to 1 no matter how large L is.

4.2 Capacity Expressions

Now consider the noisy channel, $\boldsymbol{x} = \sqrt{P}\boldsymbol{H}\boldsymbol{s} + \boldsymbol{w}$, where P is the transmitted power ($\mathrm{E}[\|\boldsymbol{s}\|^2] = 1$) and $\boldsymbol{w}$ is zero-mean complex Gaussian noise vector with $\mathrm{E}[\boldsymbol{w}\boldsymbol{w}^H] = \boldsymbol{I}$. In this section, for simplicity of exposition, we fix $n_T = n_R = n$. Given the knowledge of channel coefficients ($\boldsymbol{H}$ or $\boldsymbol{H}_V$) at the receiver, the channel capacity is given by [2, 1]

$$C(\boldsymbol{H}) = \log_2\left[\det\left(\boldsymbol{I} + P\boldsymbol{H}\boldsymbol{H}^H/n\right)\right] = \log_2\left[\det\left(\boldsymbol{I} + P\boldsymbol{H}_V\boldsymbol{H}_V^H/n\right)\right] = C(\boldsymbol{H}_V) \ bits/s/Hz \tag{15}$$

where the second equality follows from the unitary equivalence of $\boldsymbol{H}$ and $\boldsymbol{H}_V$. The ergodic capacity is given by $C_E = \mathrm{E}[C(\boldsymbol{H}_V)]$ where the expectation is over the statistics of $\boldsymbol{H}_V$. If the propagation environment consists of a cluster of scatterers, the imaging interpretation of $\boldsymbol{H}_V$ implies that it would consist of nonvanishing submatrices corresponding to different clusters. In such cases, the expression (15) can be further simplified to

$$C(\boldsymbol{H}_V) = \sum_{i=1}^{N_c} \log_2\left[\det(\boldsymbol{I} + P\boldsymbol{H}_V(i)\boldsymbol{H}_V^H(i)/n)\right] \quad bits/s/Hz \tag{16}$$

where N_c is the number of distinct clusters and $\boldsymbol{H}_V(i)$ is the submatrix corresponding to the i^{th} cluster.[3] For example, two $\pi/8$-wide clusters (as seen from either the transmitter or receiver) centered at $\phi_R = \phi_T = \pm\pi/4$ would results in a block diagonal matrix $\boldsymbol{H}_V = \text{diag}(\boldsymbol{H}_V(1), \boldsymbol{H}_V(2))$ with the size of $\boldsymbol{H}_V(i)$'s determined by the number of resolvable virtual angles within the $\pi/8$ angular spreads. Each $\boldsymbol{H}_V(i)$ in (16) can then be represented as a k-diagonal matrix with iid entries representing the nature of scattering within the cluster (the extent of cross-coupling between the transmit and receive virtual angles within the cluster). The cluster decomposition in (16) has the following simple interpretation:

An arbitrary spatial channel (scattering environment) can be decomposed into N_c independent parallel channels in the VS that are represented by the matrices $\{\boldsymbol{H}_V(i) : i = 1, \cdots, N_c\}$, each $\boldsymbol{H}_V(i)$ in turn representing a k-diagonal iid spatial channel that corresponds to the nature of scattering in each cluster.

4.3 Numerical Examples

We now illustrate capacity computation via $\boldsymbol{H}_V$ for two different environments: An idealized rich scattering environment and a more realistic scattering environment with a smaller scattering angular spread. In all examples, $n_T = n_R = 10$ and $\text{SNR} = 10\log_{10}(P) = 20\text{dB}$.

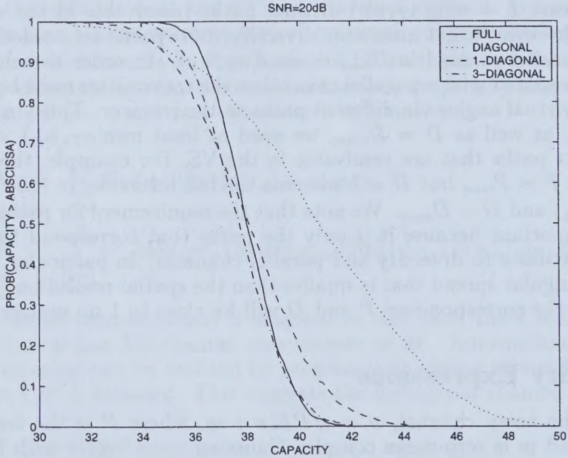


Figure 4: *Capacity comparison in a rich scattering environment with energy normalization so that all models have the same received SNR. The level of diversity, D, is the main difference between various models.*

Consider the rich scattering case. The channel in this case is directly modeled via (1) with the elements of $\boldsymbol{H}$ given by iid complex zero-mean Gaussian random variables with unit variance. We compare the capacity of a full iid model with a k-diagonal VS system. Two separate comparisons are made. In the first comparison, illustrated in Figure 4, the k-diagonal VS matrix is scaled so that the received SNR is the same as in the full iid case. The scaling factor is given by $n^2/(n + k(2n - k - 1))$, $k = 0, \cdots, n - 1$, which equals n for $k = 0$ (diagonal approximation) and 1 for $k = (n-1)$ (full matrix). This comparison captures the effect of diversity on outage capacity — a k diagonal system provides higher values of D with increasing k. As evident from Figure 4, the performance of a 3-diagonal system is very close to the full system ($k = 9$), demonstrating that for the same received SNR the 3-diagonal system captures most of the diversity advantage. We note that the ergodic capacity of the all systems is virtually identical in this case. The second comparison, illustrated in Figure 5, does not normalize the received energy can be thought of as a comparison of the full iid channel with that of a k-diagonal VS approximation (projection of the full channel onto the space of k-diagonal channels). Both the received SNR and diversity are at play in this case. As evident from Figure 5, the diagonal approximation is very poor but the 5-diagonal approximation comes within 5% of the capacity of the full iid system. This suggests that even in an idealized rich scattering environment, k-diagonal VS approximations with moderate values of k can be quite adequate. Note that k-diagonal systems result in simpler transceiver structures due to fewer channel parameters.

---

[3] The relation (16) assumes that each cluster corresponds to distinct transmit and receive virtual angles. However, this assumption can be relaxed and a relation of the form (16) can be derived in that case as well.

Figure 6 compares the capacity of a channel with k-diagonal VS approximations in the VS under a more realistic scattering environment with no SNR normalization. In this case, the channel matrix $\boldsymbol{H}$ is simulated via the physical model (2) corresponding to two $\pi/8$-wide clusters (as seen from either the transmitter or receiver) centered at $\phi_R = \phi_T = \pm\pi/4$. There are $L = 50$ paths in each cluster. The angles $\{\phi_{T,l}, \phi_{R,l}\}$ for the paths are uniformly distributed over the angular spreads and the corresponding fading gains $\{\beta_l\}$ are simulated as iid zero mean complex Gaussian random variables with unit variance. As evident from Figure 6, the capacity of a 2-diagonal system is very close to that of the full (9-diagonal) system. This is due to the fact that D is relatively small in this case due to limited angular scattering spread and is adequately captured by a 2-diagonal system (significant correlation between the elements of $\boldsymbol{H}$).

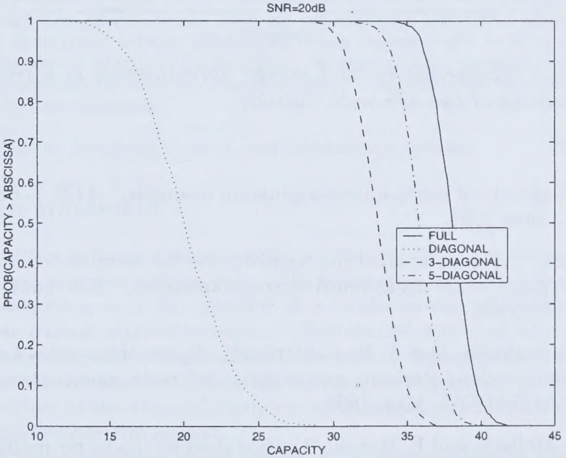


Figure 5: *Capacity comparison without energy normalization in a rich scattering environment. Both received SNR and diversity affect capacity in this case.*

5 Conclusions

We have proposed a virtual space (VS) framework for modeling multi-antenna multipath channels that provides a natural link between various existing modeling approaches. In particular, it captures physical propagation effects in a linear fashion via the notion of virtual spatial angles and provides a direct handle on computing the corresponding channel statistics. The VS channel model facilitates accurate estimates of channel capacity in realistic scattering environments. It clearly reveals the two key parameters affecting capacity: the number of parallel channels and the level of diversity per parallel channel. While the focus of this paper is on spatial effects, the framework can be readily extended to include the temporal dimension and such extensions will be reported elsewhere. Finally, the framework suggests direct extensions for modeling spatial multiplexing schemes.[4]

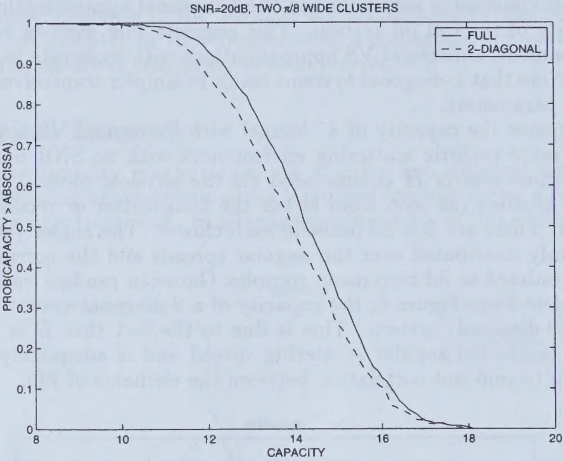


Figure 6: *Capacity comparison without energy normalization in a realistic scattering environment consisting of two $\pi/8$-wide clusters.*

References

[1] E. Telatar, “Capacity of multi-antenna gaussian channels,” *AT&T-Bell Labs Internal Tech. Memo.*, June 1995.

[2] G. J. Foschini, “Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas,” *Bell Labs Tech. J.*, vol. 1, no. 2, pp. 41–59, 1996.

[3] V. Tarokh, N. Seshadri, and A. R. Calderbank, “Space-time codes for high data rate wireless communication: Performance criteria and code construction,” *IEEE Trans. Inform. Th.*, pp. 744–765, Mar. 1998.

[4] J. Fuhl, A. F. Molisch, and E. Bonek, “Unified channel model for mobile radio systems with smart antennas,” in *IEE Proc. Radar, Sonar Navig.*, vol. 145, pp. 32–41, Feb. 1998.

[5] G. G. Raleigh and J. M. Cioffi, “Spatio-temporal coding for wireless communication,” *IEEE Trans. Commun.*, pp. 357–366, Mar. 1998.

[6] J. W. Brewer, “Kronecker products and matrix calculus in system theory,” *IEEE Trans. Circ. and Syst.*, vol. 25, pp. 772–781, Sep. 1978.

---

[4] R. W. Heath, Jr. Personal communication.

On Achivable Rates in a Multi-Antenna Broadcast Downlink

Giuseppe Caire† Shlomo Shamai‡
† Eurecom – Sophia-Antipolis – France
‡ Technion – Haifa – Israel

Abstract

A Gaussian broadcast channel with r single-antenna receivers where the transmitter is equipped with t antennas and where both the transmitter and the receivers have perfect knowledge of the propagation channel is considered. This provides a very simple model for the downlink of a wireless system, but despite its apparent simplicity it is in general a non-degraded broadcast channel, for which the capacity region is not fully known. We propose a novel transmission scheme based on "ranked known interference". In brief, the transmitter decomposes the channel into an ordered (or ranked) set of interference channels for which the interference signal of the i-th channel is generated as a linear combination of the signals transmitted in channels $j < i$. In this way, known techniques of coding for non-causally known interference can be applied to make the interference in each channel harmless without further power penalty. We show that the proposed scheme is throughputwise asymptotically optimal for both low and high SNR. In the special case of 2-antenna and 2-users we propose a modification of the basic strategy achieving optimal throughput for all SNRs. Finally, the infinite-dimensional Rayleigh channel is considered and throughput closed-form expressions are provided in various cases. This analysis shows that a practical and sensible strategy to the downlink of a wireless system consists of hybrid TDMA and space-time multiplexing where only a subset of active users, whose optimal size depends on the available SNR, is served at any given channel use by using our ranked known interference scheme. TDMA is used for time-sharing between different active user subsets to give to all users the same average rate without penalty in the maximum throughput. Also, constant-power variable-rate coding achieves practically the same throughput of variable-power variable-rate coding (waterfilling power allocation).

Keywords: Gaussian broadcast channel, multiple-antenna systems.

1 Problem statement

We consider a system with one transmitter and r receivers. The transmitter has t antennas, while the receivers have one antenna each. The transmitter has to deliver to each receiver independent information, as in the downlink of a cellular system where the base station is equipped with an antenna array of t elements.[1] This channel is referred to in the following as the $t \times 1 : r$ Gaussian Broadcast Channel (GBC), in order to stress the fact that the r receivers must process their signals separately. If, on the contrary, the receivers are allowed to cooperate, the system boils down to the standard multiple-antenna single-user channel [13]. We shall refer to this case as the "cooperative system".

Let $\mathbf{x} \in \mathbb{C}^t$ denote the vector of modulation symbols transmitted in parallel from the t antennas in a given channel use (symbol interval), let $\mathbf{H} \in \mathbb{C}^{r \times t}$ denote the channel matrix, where $h_{i,j}$ is the complex channel gain from antenna j to antenna i, let $\mathbf{y} \in \mathbb{C}^r$ denote the vector of received signal samples at the r receivers and let $\mathbf{z} \in \mathbb{C}^r$ be the corresponding vector of noise samples, assumed to be circularly-symmetric complex Gaussian with i.i.d. components such that $z_i \sim \mathcal{N}_{\mathbb{C}}(0, N_0)$. Then, the $t \times 1 : r$ GBC is described by

---

[1] More in general, the t antennas might represent the collection of all base stations, each equipped with an antenna array. However, in this work we do not consider the effect of the spatial distribution of the transmit and receive antennas as for example in the cellular model of [11].For further results along this line see [12].

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{z} \tag{1}$$

The channel matrix $\mathbf{H}$ is known to the transmitter and to all receivers and the input is constrained to satisfy $\text{trace}(\mathbf{\Sigma}_x) \leq \mathcal{E}$ where $\mathbf{\Sigma}_x \overset{\Delta}{=} E[\mathbf{x}\mathbf{x}^H]$ and $\mathcal{E}$ is the maximum allowed total transmit energy per channel use.

The $1 \times 1 : r$ GBC coincides with the classical *degraded* Gaussian broadcast channel, whose capacity region is well-known [4]. However, the $t \times 1 : r$ GBC for $t > 1$ is in general a non-degraded broadcast channel and cannot be cast in the framework of parallel Gaussian broadcast channels [5] since the receivers cannot cooperate. In this paper we are mainly concerned with the channel throughput R (or rate-sum), defined as the sum of all individual user rates for which the corresponding rate r-tuple is achievable. We shall consider the following scenarios: i) $\mathbf{H}$ is deterministic and fixed; ii) $\mathbf{H}$ is fixed during the transmission of each code word, but it is randomly selected according to a given probability distribution (composite channel).

In section 2 we outline the newly proposed transmission scheme. In section 3 we show that this scheme outperforms conventional zero-forcing (ZF) beamforming (see [1] and references therein), it is throughputwise asymptotically optimal for both low and high SNR, and in the 2-antennas 2-users case we demonstrate that a modified strategy is in fact optimal for all SNR. Finally, in section 4 we provide results for an independent Rayleigh channel in the large-system limit, i.e., when both r and t go to infinity with a given constant ratio α. The proofs of the results of this paper can be found in [2].

2 A new approach based on "ranked known interference"

In this section we present a new transmission strategy that combines linear signal processing at the transmitter with non-standard coding. The linear processing turns the original channel into an ordered set of interference channels, where the channels are given in a specific order and the interference in channel i is generated by the signals transmitted in channels $j < i$. Since all signals are generated at the same transmitting end, for each of these channels the interfering signal is known non-causally. Therefore, we can apply the results on capacity with interference known non-causally at the transmitter. Capacity and coding for Gaussian channels with Gaussian interference non-causally known at the transmitter is solved in [3], exploiting the result of [8]. More recently, a general coding technique based on *lattice precoding* that applies also to non-Gaussian interference and yields the same (optimal) result of [3] was presented in [6]. An uncoded version of this approach, where precoding is implemented as a Tomlinson-Harashima precoder [7] (a simple one-dimensional example of lattice precoding), has been independently proposed in [9].

Let $\mathbf{H} = \mathbf{G}\mathbf{Q}$ be a QR-type decomposition of $\mathbf{H}$ where $\mathbf{G} \in \mathbb{C}^{r \times m}$ is lower triangular and $\mathbf{Q} \in \mathbb{C}^{m \times t}$ has orthonormal rows (we define $m \overset{\Delta}{=} \min\{r, t\}$). The transmitted signal is obtained as $\mathbf{x} = \mathbf{Q}^H\mathbf{u}$. Let $g_{i,j}$ denote the (i, j)-th element of $\mathbf{G}$ and let $d_i = |g_{i,i}|^2$. The original channel is turned into the set of interference channels

$$y_i = g_{i,i}u_i + \sum_{j<i} g_{i,j}u_j + z_i, \quad i = 1, \ldots, m \tag{2}$$

while no information is sent to users $m+1, \ldots, r$. Since $\mathbf{Q}\mathbf{Q}^H = \mathbf{I}$, the covariance of $\mathbf{u}$ has the same trace constraint of $\mathbf{\Sigma}_x$. The signals u_i are all generated by the transmitter. Hence, they are all known non-causally by the encoder. As explained in [2], there exist coding schemes such that each user $i = 1, \ldots, m$ "sees" an interference-free channel, as if its signal $\mathbf{u}_i$ was alone. We refer to this scheme as *Ranked Known Interference* (RKI).

Proposition 1. The achievable throughput of the RKI scheme is given by

$$\left\{ \begin{array}{l} R = \sum_{i=1}^{m} \log(1 + d_i a_i) \\ \sum_{i=1}^{m} a_i = A \end{array} \right. \tag{3}$$

where $a_i \triangleq E[|u_i|^2]/N_0$ denote the SNR on the i-th channel and $A \triangleq \mathcal{E}/N_0$ denote the total *transmit* SNR. □

The RKI throughput can be optimized jointly with respect to the power allocation (the $a_i's$) and the user ordering. In fact, for an arbitrary permutation matrix $\mathbf{\Pi}$, the set of elements $\{d_i' : i = 1, \dots, m\}$ resulting from the QR decomposition $\mathbf{\Pi H} = \mathbf{G}'\mathbf{Q}'$ is generally different from the set $\{d_i : i = 1, \dots, m\}$ resulting from $\mathbf{H} = \mathbf{GQ}$.

3 Results for the deterministic channel

In this section we consider $\mathbf{H}$ given and constant. It is clear that the cooperative throughput R^{coop} (maximized w.r.t. the power allocation) is an upperbound to the throughput R^{gbc} of the $t \times 1 : r$ GBC, while the throughput of ZF and RKI, denoted by R^{zf} and by R^{rki}, respectively, provide lower bounds. The RKI scheme yields generally a larger maximal throughput than conventional ZF beamforming:

Proposition 2. For any channel matrix $\mathbf{H}$, $R^{\text{rki}-\max} \geq R^{\text{zf}-\max}$. □

The following proposition makes this statement stronger in the limits for high and low SNR:

Proposition 3. For any channel matrix $\mathbf{H}$ with full row-rank r,

$$\lim_{A \to \infty} \left(R^{\text{coop}-\max} - R^{\text{rki}-\max} \right) = 0 \tag{4}$$

For any channel matrix $\mathbf{H}$,

$$\lim_{A \to 0} \frac{R^{\text{rki}-\max}}{R^{\text{zf}-\max}} = 1 \tag{5}$$

□

As a corollary of Proposition 2, we get that (for $\mathbf{H}$ of rank r) the RKI is asymptotically throughputwise optimal for large SNR, since the inequality $R^{\text{rki}-\max} \leq R^{\text{gbc}} \leq R^{\text{coop}-\max}$ and the first limit of Proposition 2 imply that $\lim_{A \to \infty}(R^{\text{gbc}} - R^{\text{rki}-\max}) = 0$.

Next, we find two upperbounds to R^{gbc} tighter than the trivial maximum cooperative throughput bound. The capacity region of a general broadcast channel depends only on the transition marginal probability assignments $p(y_i|\mathbf{x})$ and not on the whole joint transition probability $p(y_1, \dots, y_r|\mathbf{x})$ [10, 4]. For the $t \times 1 : r$ GBC, this implies that the channels in the family defined by

$$\mathbf{y} = \mathbf{\Pi H x} + \mathbf{z} \tag{6}$$

where $\mathbf{H}$ is given, $\mathbf{\Pi}$ is any $r \times r$ permutation matrix, $\text{trace}(\mathbf{\Sigma}_x) \leq A$ and $\mathbf{z} \sim \mathcal{N}_{\mathbb{C}}(\mathbf{0}, \mathbf{\Sigma}_z)$ where $\mathbf{\Sigma}_z$ is any non-negative definite Hermitian matrix with diagonal elements equal to 1 (we refer to this constraint as the *unit-diagonal constraint*), have all the same broadcast capacity region (and therefore the same R^{gbc}). Hence, a tighter cooperative throughput upperbound is obtained by choosing the worst-case (cooperative throughputwise) channel in the family defined by (6). We have:

Lemma 1. For any channel matrix $\mathbf{H}$,

$$R^{\text{gbc}} \leq \max_{\mathbf{\Sigma}_x} \min_{\mathbf{\Pi}, \mathbf{\Sigma}_z} \left\{ \log\det\left(\mathbf{I} + \mathbf{\Sigma}_z^{-1} \mathbf{\Pi H \Sigma}_x \mathbf{H}^H \mathbf{\Pi}^H \right) \right\} \tag{7}$$

where minimization is over all $r \times r$ permutation matrices $\mathbf{\Pi}$ and over all noise covariances $\mathbf{\Sigma}_z$ with unit diagonal elements and maximization is over all input covariances with trace not larger than A. ☐

Assume that $\mathbf{H}$ has rank $k \leq m$ and, after a suitable row permutation, can be partitioned as $\mathbf{H} = [\mathbf{H}_1^T, \mathbf{H}_2^T]^T$ where $\mathbf{H}_1 \in \mathbb{C}^{k \times t}$ has rank k and $\mathbf{H}_2 \in \mathbb{C}^{(r-k)\times t}$. Then, any row of $\mathbf{H}_2$ can be expressed as a linear combination of rows of $\mathbf{H}_1$, i.e., we can write $\mathbf{H}_2 = \mathbf{B}\mathbf{H}_1$ where $\mathbf{B} = \mathbf{H}_2\mathbf{H}_1^+$. Moreover, partition the noise vector as $\mathbf{z} = [\mathbf{z}_1^T, \mathbf{z}_2^T]^T$, and let $\mathbf{\Sigma}_{z1}$ and $\mathbf{\Sigma}_{z2}$ denote the $k \times k$ upper left and $(r-k) \times (r-k)$ lower right diagonal blocks of $\mathbf{\Sigma}_z$ (notice that both $\mathbf{\Sigma}_{z1}$ and $\mathbf{\Sigma}_{z2}$ must satisfy the unit diagonal constraint). Then, we have the following:

Lemma 2. With the above notation, if $\mathbf{\Sigma}_{z2} - \mathbf{B}\mathbf{\Sigma}_{z1}\mathbf{B}^H$ is non-negative definite, then

$$R^{\text{gbc}} \leq I(\mathbf{x}; \mathbf{y}_1) \tag{8}$$

where $\mathbf{y}_1 = \mathbf{H}_1\mathbf{x} + \mathbf{z}_1$. ☐

The above lemma can be applied to the simple choice $\mathbf{\Sigma}_{z1} = \mathbf{I}_k$ and $\mathbf{\Sigma}_{z2} = \mathbf{I}_{r-k}$. If $\|\mathbf{B}\|_2 = \sqrt{\rho(\mathbf{B}\mathbf{B}^H)} \leq 1$, then R^{gbc} is not larger than $R^{\text{coop}-\max}$ of the $t \times k$ channel defined by the matrix $\mathbf{H}_1$ (with i.i.d. Gaussian noise). From Proposition 3 we get that this throughput is asymptotically achieved for large SNR by the RKI scheme (in fact, $\mathbf{H}_1$ has full row-rank k). Then, as a consequence of Proposition 3 and Lemma 2, we have the following:

Corollary 1. Let $\mathbf{H}$ have rank $k \leq r$ and, after a suitable row permutation, assume that

$$\mathbf{H} = \begin{bmatrix} \mathbf{I}_k \\ \mathbf{B} \end{bmatrix} \mathbf{H}_1 \tag{9}$$

where $\mathbf{H}_1 \in \mathbb{C}^{k \times t}$ has rank k and $\|\mathbf{B}\|_2 \leq 1$. Then, $\lim_{A \to \infty}(R^{\text{gbc}} - R^{\text{rki}-\max}) = 0$. ☐

The following linear-algebra lemma (new to the autors' knowledge) shows that $\mathbf{H}_1$ is essentially unique up to row permutations:

Lemma 3. Let $\mathbf{H} \in \mathbb{C}^{r \times t}$ and assume that it can be decomposed as in Corollary 1. Let $\mathbf{H}_1' \in \mathbb{C}^{k \times t}$ and $\mathbf{H}_2' \in \mathbb{C}^{(r-k)\times t}$ be obtained by exchanging some rows of $\mathbf{H}_1$ with some rows of $\mathbf{H}_2 = \mathbf{B}\mathbf{H}_1$, such that that $\mathbf{H}_1'$ has rank k. Then, $\mathbf{B}' = \mathbf{H}_2'(\mathbf{H}_1')^+$ has 2-norm ≥ 1. ☐

Unfortunately, the partition of Corollary 1 is not always possible (it is immediate to find a counterexample for $r \geq 3$). Then, we cannot exclude the possibility that for such type of rank-deficient channel matrices $\mathbf{H}$ the RKI scheme is asymptotically suboptimal for large SNR.

3.1 The 2-antennas 2-users case

To fix the ideas and demonstrate results and techniques we consider in the details the example of $t = r = 2$. We have the following:

Proposition 4. For any 2×2 channel matrix $\mathbf{H}$,

$$\lim_{A \to 0} \frac{R^{\text{gbc}}}{R^{\text{rki}-\max}} = 1 \tag{10}$$

☐

The above proposition is proved by applying Lemma 1 where $\mathbf{\Pi}$ is the permutation that puts in first position the row of $\mathbf{H}$ with the largest 2-norm and with the choice

$$\mathbf{\Sigma}_z = \begin{bmatrix} 1 & \frac{g_{2,1}^*}{g_{1,1}^*} \\ \frac{g_{2,1}}{g_{1,1}} & 1 \end{bmatrix}$$

Since by construction $|g_{1,1}|^2 \geq |g_{2,2}|^2 + |g_{2,1}|^2$, this is a valid covariance matrix. In the case where the above inequality is strict the statement of Proposition 4 is actually stronger, since for $A \in [0, A_1]$, where A_1 is given by

$$A_1 = \frac{1}{|g_{1,1}|^2}\left(\frac{|g_{1,1}|^2 - |g_{2,1}|^2}{|g_{2,2}|^2} - 1\right) \tag{11}$$

the RKI scheme is (not only asymptotically) optimal.

As a corollary of Proposition 4 and Proposition 3, we have that the RKI and the ZF schemes are throughput-wise asymptotically optimal for the $2 \times 1 : 2$ GBC and low SNR. Moreover, since if $\mathbf{H}$ has rank 1 the partition (9) of Corollary 1 is always possible, the RKI scheme is asymptotically optimal for both high and low SNR for any $\mathbf{H}$. Because of these nice properties, we might be tempted to conjecture that the RKI scheme is actually optimal for all SNRs. The following modified strategy shows that this is actually not the case.

An optimal modified RKI strategy for the $2 \times 1 : 2$ GBC. We let $\mathbf{H} = \mathbf{GQ}$ and construct the transmitted signal as $\mathbf{x} = \mathbf{Q}^H\mathbf{Ru}$, where $\mathbf{R}$ is upper triangular. The resulting two channels are

$$\begin{aligned} y_1 &= g_{1,1}r_{1,1}u_1 + g_{1,1}r_{1,2}u_2 + z_1 \\ y_2 &= (g_{2,1}r_{1,2} + g_{2,2}r_{2,2})u_2 + g_{2,1}r_{1,1}u_1 + z_2 \end{aligned} \tag{12}$$

The signals u_1 and u_2 carry information for user 1 and 2, respectively, and the signal u_2 is constructed according the "coding for known interference" method (see [2]), by treating $g_{2,1}r_{1,1}u_1$ as known interference. All signals are Gaussian, and u_1, u_2 are asymptotically uncorrelated for large block length. Receiver 1 treats the signal $g_{1,1}r_{1,2}u_2$ as background noise, as this is *unknown* Gaussian interference (worst-case noise for Gaussian input). Receiver 2 is able to reliably decode the same rate as if the interference signal $g_{2,1}r_{1,1}u_1$ was not present. Therefore, the resulting throughput is given by

$$R^{\text{mod}} = \log\left(1 + \frac{|g_{1,1}r_{1,1}|^2 a_1}{1 + |g_{1,1}r_{1,2}|^2 a_2}\right) + \log\left(1 + |g_{2,1}r_{1,2} + g_{2,2}r_{2,2}|^2 a_2\right) \tag{13}$$

(where "mod" stands for *modified RKI*) subject to the constraint

$$|r_{1,1}|^2 a_1 + (|r_{1,2}|^2 + |r_{2,2}|^2) a_2 = A$$

The above throughput can be maximized with respect to a_1, a_2 and the coefficients $r_{1,1}, r_{1,2}, r_{2,2}$. We reparameterize the problem by letting $b = |g_{2,1}/g_{2,2}|$, $z = |r_{2,2}/r_{1,2}|$, $p(z) = (b+z)^2/(1+z^2)$, $q(z) = 1/(1+z^2)$, $X_1 = |r_{1,1}|^2 a_1$ and $X_2 = (|r_{1,2}|^2 + |r_{2,2}|^2) a_2$. After some algebra, we can write

$$R^{\text{mod}} = \log\left(1 + |g_{2,2}|^2 p(z) X_2 + \frac{1 + |g_{2,2}|^2 p(z) X_2}{1 + |g_{1,1}|^2 q(z) X_2} |g_{1,1}|^2 X_1\right) \tag{14}$$

with the constraint $X_1 + X_2 = A$. Obviously, since $\lim_{z\to\infty} p(z) = 1$ and $\lim_{z\to\infty} q(z) = 0$, the modified RKI strategy coincides with standard RKI for $z \to \infty$. Therefore, this is in fact a *generalization* of RKI and cannot yield worse results after maximizing with respect to the power allocation X_2, X_1 and the free parameter $z \in \mathbb{R}_+$.

By substituting $X_1 = A - X_2$ in (14), R^{mod} can be maximized with respect to X_2 over the interval $[0, A]$ for any value of z. The equation $\frac{\partial}{\partial X_2} R^{\text{mod}} = 0$ has either none or one solution in the interval $[0, A]$. It can be shown that if there is no solution, the derivative is negative and the maximum is achieved by $X_2 = 0$ while if there is a solution it corresponds to a maximum. The case $X_2 = 0$ yields $R^{\text{mod}} = \log(1 + |g_{1,1}|^2 A)$, which coincides with the maximum throughput of RKI and ZF in the range $A \in [0, A_1]$ where A_1 is given by (11). For $A > A_1$, there exists a range of $z \in \mathbb{R}_+$ for which the maximum is achieved by $0 < X_2 < A$. In this range of SNR, the modified RKI strategy is distinctly better than standard RKI. Unfortunately, the maximization with respect to z must be performed numerically.

With some more effort, we are able to prove that the modified RKI strategy is actually throughputwise optimal for all SNR. Remarkably, this is one of the rare examples of a non-degraded Gaussian broadcast channel for which the optimal throughput is known. The direct part of this statement is represented by the modified RKI strategy outlined above. The converse part is provided by the upperbound of Lemma 1. First, we choose the permutation $\mathbf{\Pi}$ such that $\mathbf{\Pi}\mathbf{H} = \mathbf{G}\mathbf{Q}$ with $|g_{1,1}|^2 \geq |g_{2,2}|^2 + |g_{2,1}|^2$, as in the proof of Proposition 4. Then, we notice that any signal covariance matrix can be parameterized as

$$\mathbf{\Sigma}_x = \mathbf{Q}^H \mathbf{R} \mathbf{\Lambda}_x \mathbf{R}^H \mathbf{Q}$$

where $\mathbf{\Lambda}_x = \text{diag}(P_1, P_2)$ and where $\mathbf{R}$ is upper triangular, while any noise covariance matrix satisfying the unit diagonal constraint can be written as

$$\mathbf{\Sigma}_z = \left[\begin{array}{cc} 1 & \rho \\ \rho^* & 1 \end{array} \right]$$

where $|\rho|^2 \leq 1$. After some algebra, the upper bound of Lemma 1 can be written as

$$\begin{aligned} R^{\text{gbc}} \leq & \max_{z \geq 0, X_1 + X_2 = A} \min_{|\rho| \leq 1} \log \Bigg[1 + \frac{1}{1 - |\rho|^2} \big(|g_{2,2}|^2 p(z) X_2 + \\ & + (|g_{1,1}|^2 - 2|g_{2,1}||g_{1,1}||\rho| + |g_{2,1}|^2)(X_1 + q(z)X_2) + \\ & + X_2 q(z)(z^2 |g_{1,1}|^2 |g_{2,2}|^2 X_1 - 2|g_{1,1}||g_{2,2}||\rho| z - |g_{2,1}|^2) \big) \Bigg] \end{aligned} \qquad (15)$$

where $p(z)$ and $q(z)$ have been defined previously. The minimization with respect to $|\rho| \in [0, 1]$ yields $|\rho| = \frac{1}{2}\left(B - \sqrt{B^2 - 4}\right)$ where

$$B = \frac{(|g_{2,1}|^2 + |g_{1,1}|^2)X_1 + (|g_{2,2}|^2 p(z) + |g_{1,1}|^2 q(z))X_2 + q(z) z^2 |g_{1,1}|^2 |g_{2,2}|^2 X_1 X_2}{|g_{2,1}||g_{1,1}|(X_1 + q(z)X_2) + q(z) z |g_{1,1}||g_{2,2}| X_2}$$

is always ≥ 2 for all $z \in \mathbb{R}_+$ and $g_{1,1}, g_{2,1}, g_{2,2}$ such that $|g_{1,1}|^2 \geq |g_{2,2}|^2 + |g_{2,1}|^2$. Notice that for $X_2 = 0$ we get $|\rho| = |g_{2,1}/g_{1,1}|$, i.e., in the range $A \in [0, A_1]$ the bound coincides (obviously!) with that of Proposition 4. After substituting the above solution for $|\rho|$ and $X_1 = A - X_2$ into (15), the maximization with respect to X_2 and z must be carried out numerically. Interestingly, the maximum is achieved for the same pair (z, X_2) that maximizes R^{mod}, even though the upper bound does not coincides with R^{mod} for all (z, X_2).

Example 1. Fig. 1 shows $R^{\text{rki-max}}$, $R^{\text{mod-max}}$, $R^{\text{coop-max}}$, the upperbound to R^{gbc} given by Proposition 4 and the upperbound obtained by the above application of Lemma 1 (denoted by RKI, MOD, COOP, UB1 and UB2, respectively) for a channel with $\mathbf{G}$ factor

$$\mathbf{G} = \left[\begin{array}{cc} 1 & 0 \\ 0.9 & 0.2 \end{array} \right]$$

$R^{\text{mod-max}}$ and the upperbound UB2 coincide exactly for all SNR (thicker line). Notice that $R^{\text{rki-max}}$ is optimal in the range $A \in [0, A_1]$ and strictly suboptimal for larger A. ◊

4 Results for the infinite-dimensional Rayleigh channel

In this section we focus on the composite channel when the channel matrix has i.i.d. entries $\sim \mathcal{N}_{\mathbb{C}}(0, 1)$, and we consider the throughput achievable by the basic RKI strategy under various power constraints, assuming that no effort is made to optimize the user ordering. For finite t and r, closed-form expressions for the throughput are obtained in [2]. Here we focus on the large-system limits, i.e., we let $r \to \infty$ with $r/t = \alpha$, where $\alpha \geq 0$ is fixed and represents the channel *spatial* load expressed in "users per antenna".

In the composite channel setting, the channel is symmetric with respect to any user. Therefore, if a given ergodic throughput R is achievable for a given set of active users, by time-sharing with uniform probability over all possible user subsets (and orderings) every user can achieve the same per-user average rate R/r. As an effect of this "ergodic symmetrization", requiring equal average per-user rates involves no loss of optimality in the overall throughput. Moreover, as $r \to \infty$ the instantaneous throughput converges almost surely to the ergodic (or average) throughput [15, 14]. We let $\rho \triangleq \frac{1}{r}R$ denote the normalized throughput, equal to the symmetric per-user rate achievable by time-sharing argument outlined above. It can be shown (see [2] for the details) that in the limit of large system the RKI normalized throughput is given by

$$\left\{ \begin{array}{l} \rho^{\rm rki} = \int_0^\mu \log\left(1 + [1 - \alpha\nu]_+ a(\nu)\right) \, d\nu \\ \int_0^\mu a(\nu) \, d\nu = A/\alpha \end{array} \right. \tag{16}$$

where $a(\nu)$ is the transmit SNR of the $\lfloor \nu r \rfloor$-th signal and $\mu \triangleq m/r = \min\{1, 1/\alpha\}$.

Uniform powers. Let $\kappa = k/r$ denote the fraction of active users. In general, $\kappa \in [0, \mu]$ since from Lemma 5 no more than $\lfloor \mu r \rfloor$ users can be served in parallel. For a given κ, we consider the uniform power allocation $a(\nu) = a$ for $\nu \leq \kappa$ and $a(\nu) = 0$ for $\nu > \kappa$. We obtain

$$\begin{aligned} \rho^{\rm rki-eq} &= \frac{1}{A}\left[(\kappa + A/\alpha)\log(\kappa + A/\alpha) - \right. \\ &\quad \left. - (\kappa + (1 - \alpha\kappa)A/\alpha)\log(\kappa + (1 - \alpha\kappa)A/\alpha)\right] - \kappa(\log(\kappa) - 1) \end{aligned} \tag{17}$$

This can be further maximized with respect to $\kappa \in [0, \mu]$.

Equal rates. Again, let $\kappa < \mu$ be the fraction of active users that must be served with equal rate $R_u = \log(1 + a_0)$. The resulting normalized throughput is given by $\rho^{\rm rki-cr} = \kappa R_u$ and where a_0 is the solution of

$$\int_0^\kappa \frac{a_0}{1 - \alpha\nu} d\nu = A/\alpha \tag{18}$$

yielding $a_0 = -A/\log(1 - \alpha\kappa)$. The resulting normalized throughput is given by

$$\rho^{\rm rki-cr} = \kappa \log\left(1 - \frac{A}{\log(1 - \alpha\kappa)}\right)$$

and can be further maximized with respect to $\kappa \in [0, \mu)$.

Maximum throughput. In this case, we want to maximize $\rho^{\rm rki}$ subject to the input constraint. The standard waterfilling solution yields

$$\rho^{\rm rki-max} = \left\{ \begin{array}{ll} \log\left[\frac{1}{\alpha}(A - \log(1 - \alpha))\right] - (1/\alpha - 1)\log(1 - \alpha) - 1 & \text{(case 1)} \\ \frac{1}{\alpha}\left(\log\xi + \frac{1}{\xi} - 1\right) & \text{(case 2)} \end{array} \right. \tag{19}$$

where "case 1" corresponds to the condition $\{\alpha < 1, A \geq \frac{\alpha}{1-\alpha} + \log(1 - \alpha)\}$ and "case 2" to the complement condition and where ξ in case 2 is the solution of $\xi - \log\xi - 1 = A$.

ZF and cooperative throughput. For the sake of comparison, we calculate also the normalized throughput with ZF beamforming and with cooperative receivers in the large-system regime. We omit the details for the sake of space limitation (see [2] and references therein).

Results. Fig.2 shows the normalized throughput (or per-user rate) for all the cases discussed above versus the transmit SNR A, for $\alpha = 1.0$. In all cases, the throughput obtained by uniform power allocation with optimization of the fraction of active users (denoted by "EQ") is almost identical to that obtained by the optimal waterfilling power allocation (denoted by "MAX"). This is in accordance with the well-known fact of standard ISI channels, for which optimizing the transmission bandwidth and with a rectangular input power spectral density buys almost all the capacity achievable by waterfilling. RKI with constant per-user rate (denoted by "CR") performs slightly worse than with uniform (or waterfilling) power allocation. Finally, ZF performs quite worse than RKI (in all cases) even after optimizing the fraction of active users. Fig. 3 shows the optimal fraction of active users κ versus A for RKI-EQ, RKI-CR and ZF. The fraction of active users is an increasing function of the input SNR.

From these examples, we argue that a practical and sensible downlink system design consists of applying the RKI scheme by transmitting with constant-power variable-rate user codes and by selecting carefully the number of active users k according to the available transmit average power. Even if $r > k$ users are to be served, the system should allow only k active per channel use (say, in each slot), where k is optimally selected, and serve all r users equally by time-sharing. Therefore, the system works according to a hybrid TDMA (time-sharing) and "space-time multiplexing" given by the RKI scheme.

5 Conclusions

We proposed a new coding scheme based on *Ranked Known Interference* for the Gaussian broadcast channel with multiple antennas at the transmitter. The basic RKI scheme is asymptotically throughputwise optimal for both low and high SNR, and can be approximated in practice by some form of lattice precoding. We also exhibited a modified RKI scheme which is indeed optimal for all SNRs in the 2-users 2-antennas case. The modified RKI scheme can be applied for any t and r and might provide a handle to the study of individual achievable rates for the general $t \times 1 : r$ GBC, though there is little hope that this mechanism can determine the full capacity region.

References

[1] P. Baier, M. Meurer, T. Weber, and H. Troeger. Joint transmission (JT), an alternative rationale for the downlink of time division CDMA using multi-element transmit antennas. In *IEEE 6th Int. Symp. on Spread-Spectrum Tech. & Appl. (ISSSTA)*, pages 1–5, NJIT, New Jersey, USA, September 2000.

[2] G. Caire and S. Shamai. On achievable rates in a multi-antenna broadcast downlink. In preparation, 2000.

[3] M. Costa. Writing on dirty paper. *IEEE Trans. on Inform. Theory*, 29(3):439–441, May 1983.

[4] T. Cover and J. Thomas. *Elements of information theory*. Wiley, New York, 1991.

[5] D.Tse. Optimal power allocation over parallel Gaussian broadcast channels. *IEEE Trans. on Inform. Theory*, to appear.

[6] U. Erez, S. Shamai, and R. Zamir. Capacity and lattice-strategies for cancelling known interference. In *ISITA 2000*, Honolulu, Hawaii, USA, November 2000.

[7] V. Eyuboglu and D. Forney. Combined equalization and coding using precoding. *IEEE Comm. Magazine*, December 1991.

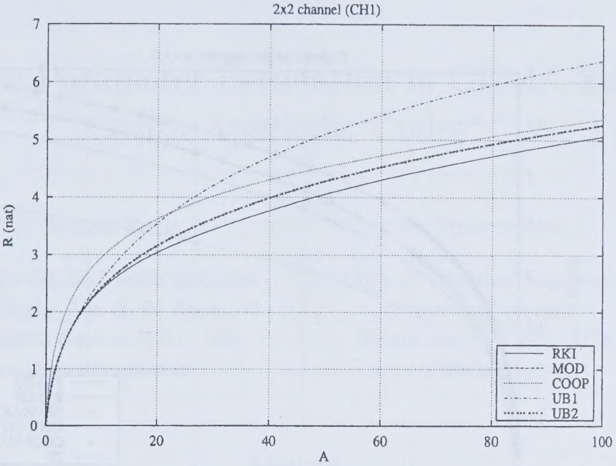


Figure 1: Channel matrix CH1: throughput (in nat) vs. A (not in dB) for RKI, cooperative (COOP), for the modified RKI scheme (MOD) and for the upperbound of Proposition 4 (UB1) and the upperbound of Lemma 1 (UB2).

[8] S. Gelfand and M. Pinsker. Coding for channel with random parameters. *Problems of Control and Information Theory*, 9(1):19–31, January 1980.

[9] G. Ginis and J. Cioffi. Vectored-DMT: a FEXT canceling modulation scheme for coordinating users. Submitted to ICC 2001, October 2000.

[10] H. Sato. An outer bound on the capacity region of broadcast channels. *IEEE Trans. on Inform. Theory*, 24(3):374–377, May 1978.

[11] S. Shamai and A. Wyner. Information theoretic considerations for symmetric cellular, multiple-access fading channels – Part I and II. *IEEE Trans. on Inform. Theory*, 43(6):1877–1911, November 1997.

[12] S. Shamai, B. Zaidel, and O. Zeitouni. Enhancing the cellular downlink capacity via co-processing at the transmitting end. In *VTC 2001*, Tel-Aviv, Israel, May 2001.

[13] E. Telatar. Capacity of multi-antenna Gaussian channels. *European Trans. on Telecomm. ETT*, 10(6):585–596, November 1999.

[14] S. Verdu and S. Shamai. Spectral efficiency of CDMA with random spreading. *IEEE Trans. on Inform. Theory*, 45(2):622–640, March 1999.

[15] S. Verdu and S. Shamai. The effect of frequency-flat fading on the spectral efficiency of CDMA. to appear on IEEE Trans. on Inform. Theory, 2000.

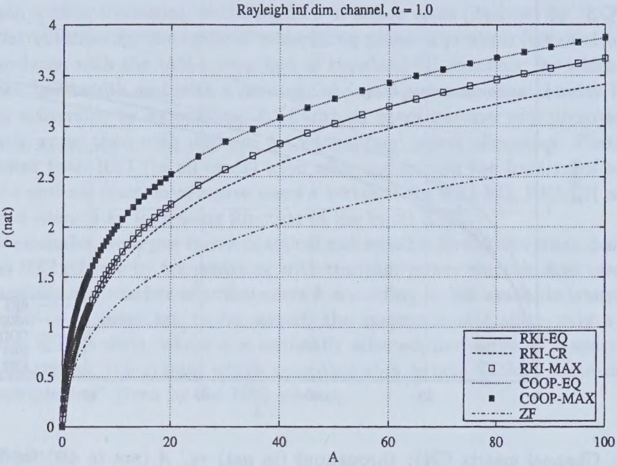


Figure 2: Normalized throughput versus transmit SNR for the Rayleigh infinite-dimensional channel with $\alpha = 1.0$.

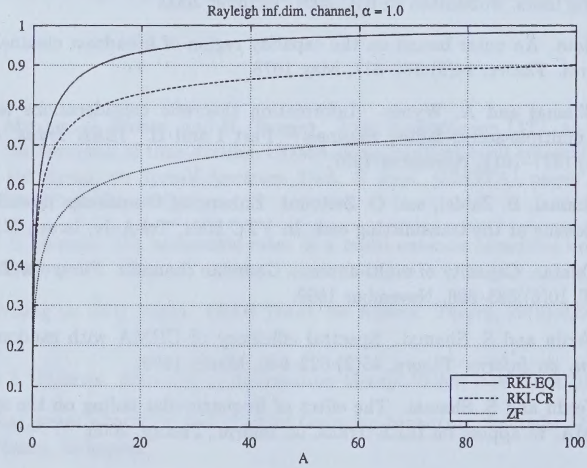


Figure 3: Optimal fraction of active users versus transmit SNR for the Rayleigh infinite-dimensional channel with $\alpha = 1.0$.

Adaptive Parameter Estimation in CDMA Systems with Aperiodic Codes*

Stefano Buzzi
Universitá degli Studi di Cassino
DAEIMI - Via G. Di Biasio, 43
I-03043 Cassino (FR) - Italy
buzzi@unicas.it

H. Vincent Poor
Dept. of Electrical Engineering
Princeton University
Princeton, NJ 08544 - USA
poor@princeton.edu

Abstract

The problem of parameter estimation in DS/CDMA systems employing long (aperiodic) spreading codes is considered in this paper. In particular, we focus on an asynchronous network and consider the problem of estimating the amplitudes, phase offsets, propagation delays and directions of arrival for the CDMA signals transmitted by the active users. The procedure that we propose belongs to the family of least-squares algorithms: it is based on the transmission of known pilot symbols, and enables adaptive estimation of the said parameters with a computational complexity which is only quadratic in the processing gain. Finally, some computer simulation results illustrating the performance of the newly proposed estimation algorithms are provided and discussed.

1 Introduction

A key point in the design of a multiuser receiver is the synthesis of high-performance parameter estimation algorithms. Indeed, any multiuser detector relies on the knowledge of several parameters, like, for example, the amplitude, the relative delay, the direction-of-arrival (DoA), and the phase offset of the CDMA signals. While the most part of the papers addressing the problem of multiuser detection assume that these parameters are known, in practical systems they are to be estimated prior to data demodulation through a suitable processing of the received waveform. Classical parameter estimation methods [1], which assume that the additive noise is a sample function of a complex zero-mean Gaussian random process, exhibit a very poor performance in multiuser environments, especially in a near-far scenario, and indeed the problem of parameter estimation in the presence of Gaussian thermal noise and multiaccess interference (MAI) is an active research topic [2, 3, 4, 5, 6].

Despite such an huge amount of existing results, it should be noted that the most part of the existing algorithms is based on the *crucial* hypothesis that the DS/CDMA system is employing short (periodic) codes. On the other hand, even though short codes

---

*This work was supported partly by the Italian National Reseach Council (CNR) under a NATO/CNR Grant, and partly by the U. S. National Science Foundation under Grant CCR-99-79361.

are an option, the majority of the third-generation CDMA-based wireless networks will be adopting long (aperiodic) spreading sequences [7, 8]; likewise, the second-generation CDMA wireless network standard IS-95 employs long codes too. Since long-codes destroy the bit-interval cyclostationarity properties of the CDMA signals, many of the parameter estimation procedures so far proposed in the open literature (like, for instance, the MUSIC-type algorithms) cannot be applied to systems with aperiodic long codes; some other algorithms, on the contrary, may be still adopted in long-code CDMA systems, but with an unaffordable computational complexity increase.

In this paper, we thus deal with the problem of parameter estimation in CDMA systems with aperiodic codes. In particular, we consider the case that the receiver is equipped with an antenna array and dwell on the problem of joint estimation of the following signal parameters: amplitude, phase offset, propagation delay and Direction of Arrival (DoA). We derive estimation procedures for both the cases that either the parameters from all of the active users are to be jointly estimated (multiuser estimation) or the receiver is interested in estimating the parameters of the signal transmitted by one user only (single-user estimation). We also provide some computer simulations results so as to show the effectiveness of the newly proposed estimation procedures. The paper is organized as follows. In the next Section we derive the received signal model for an asynchronous DS/CDMA system employing long codes. In Section 3 we illustrate the adaptive estimation algorithms for single-user and multiuser CDMA signals parameter estimation, while numerical results and computer simulations examples are given in Section 4. Finally, concluding remarks are drawn in Section 5.

Notation: In the following, the superscripts $(\cdot)^*$, $(\cdot)^T$ and $(\cdot)^H$ denote conjugate, transpose and conjugate transpose, respectively; $E[\cdot]$ denotes statistical expectation; the symbols $\otimes$ and $\odot$ denote the Kronecker product and the Schur (i.e. component-wise) matrix product, respectively; column-vectors and matrices are indicated through boldface lowercase and uppercase letters, respectively. The terms $\boldsymbol{I}_s$ and $\boldsymbol{O}_s$ denote the identity and all-zero matrix of order s, respectively; finally, $\text{Diag}(\boldsymbol{a})$ denotes a diagonal matrix containing the entries of the vector $\boldsymbol{a}$ on its diagonal.

2 System Model

Let us consider an asynchronous DS/CDMA System employing aperiodic long-codes. Denoting by K the number of active users, and assuming that the receiver is equipped with an uniform linear antenna array with Q elements, the baseband equivalent of the received signal at the q-th antenna is written as:

$$r_q(t) = \sum_{p=-P}^{P} \sum_{k=0}^{K-1} \alpha_{k,q} A_k e^{j\phi_k} b_k(p) s_{k,p}(t - \tau_k - pT_b) + n_q(t) \quad q = 0, \ldots, Q-1 \qquad (1)$$

In the above expression, $(2P+1)$ is the transmitted packet length, T_b is the bit-interval duration, $A_k e^{i\phi_k}$ and τ_k denote the complex amplitude and timing offset of the k-th user, $b_k(p) \in \{+1, -1\}$ is the k-th user information bit in the p-th signaling interval, while $n_q(t)$ is the complex envelope of the additive noise term at the q-th antenna, that we assume to be a zero-mean, Wide-Sense Stationary (WSS) complex white Gaussian process with Power Spectral Density (PSD) $2\mathcal{N}_0$. As to $\alpha_{k,q}$, it is a complex term depending on the

DoA of the signal transmitted by the k-th user; its expression is given by:

$$\alpha_{k,q} = e^{j\frac{2\pi d(q-1)}{\lambda}\sin(\theta_k)} \tag{2}$$

with θ_k the k-th user DoA, d the spacing between two consecutive antenna elements and λ the wavelength corresponding to the signal carrier frequency. Finally, $s_{k,p}(t)$ is the k-th user signature waveform in the p-th signaling interval: observe that such a quantity, which is supported on the interval $[\tau_k, \tau_k + T_b]$, explicitly depends on the temporal index p since we are considering a DS/CDMA system with aperiodic spreading. In particular, in a DS system the signature waveform is written as

$$s_{k,p}(t) = \sum_{n=0}^{N-1} c_{k,p}(n) u_{T_c}(t - nT_c) \tag{3}$$

In (3), N is the processing gain, $T_c = T_b/N$ is the chip interval, $\{c_{kp}(n)\}_{n=0}^{N-1}$ is the spreading sequence for the k-th user in the p-th signaling interval and $u_{T_c}(\cdot)$ is a unit-height rectangular waveform of duration T_c.

In order to deal with a discrete-time signal, the waveform $r_q(t)$ is fed to a low-pass filter whose impulse response is an unit-energy rectangular waveform of duration T_c/M, with M an integer number; the output of such a filter is then sampled at rate M/T_c.

Now, if we let:

$$i_k = \left\lfloor \frac{M\tau_k}{T_c} \right\rfloor \qquad \delta_k = \tau_k - \frac{i_k T_c}{M} \qquad k = 0, \ldots, K-1 \tag{4}$$

and define the following $2NM$-dimensional vectors:

$$\boldsymbol{c}_{k,p}^{(\ell)} = \Big[\underbrace{0,\ldots,0}_{\ell}, \underbrace{c_{k,p}(0),\ldots,c_{k,p}(0)}_{M}, \ldots, \underbrace{c_{k,p}(N-1),\ldots,c_{k,p}(N-1)}_{M}, \underbrace{0,\ldots,0}_{NM-\ell}\Big]^T \tag{5}$$

$$\boldsymbol{s}_{k,p} = (1 - \delta_k M/T_c)\boldsymbol{c}_{k,p}^{(i_k)} + (\delta_k M/T_c)\boldsymbol{c}_{k,p}^{(i_k+1)} \tag{6}$$

it can be easily shown that the discrete-time version of the waveform $r_q(t)$, as observed in the p-th signaling interval $[pT_b, (p+1)T_b]$, is represented by the following NM-dimensional vector:

$$\boldsymbol{r}_q(p) = \sum_{k=0}^{K-1} \alpha_{k,q} A_k e^{j\phi_k} \left(b_k(p-1)\boldsymbol{s}_{k,p-1}^l + b_k(p)\boldsymbol{s}_{k,p}^u\right) + \boldsymbol{n}_q(p) \quad q = 0, \ldots, Q-1 \tag{7}$$

In the above equation, the NM-dimensional vectors $\boldsymbol{s}_{k,p-1}^l$ and $\boldsymbol{s}_{k,p}^u$ are obtained by retaining the last (i.e., lowest) and the first (i.e., uppest) NM entries of the $2NM$-dimensional vectors $\boldsymbol{s}_{k,p-1}$ and $\boldsymbol{s}_{k,p}$, respectively, i.e.:

$$\boldsymbol{s}_{k,p-1}^l = [\boldsymbol{O}_{NM} \;\; \boldsymbol{I}_{NM}]\, \boldsymbol{s}_{k,p-1} \qquad \boldsymbol{s}_{k,p}^u = [\boldsymbol{I}_{NM} \;\; \boldsymbol{O}_{NM}]\, \boldsymbol{s}_{k,p} \tag{8}$$

while $\boldsymbol{n}_q(p)$ represents the thermal noise contribution on the q-th antenna and is a random vector whose entries are independent complex zero-mean Gaussian random variates with variance $2\mathcal{N}_0$.

3 Adaptive Parameter Estimation

In this section, we derive pilot-assisted adaptive algorithms aimed at estimating the CDMA signals parameters $\boldsymbol{\tau} = [\tau_0, \dots, \tau_{K-1}]^T$, $\boldsymbol{\theta} = [\theta_0, \dots, \theta_{K-1}]^T$, $\boldsymbol{\Phi} = [\phi_0, \dots, \phi_{K-1}]^T$ and $\boldsymbol{a} = [A_0, \dots, A_{K-1}]^T$.

To begin with, let us define the following $2NM \times (NM+1)$-dimensional matrix:

$$\boldsymbol{C}_{k,p} = \left[\boldsymbol{c}_{k,p}^{(0)} \; \boldsymbol{c}_{k,p}^{(1)} \dots \boldsymbol{c}_{k,p}^{(NM)}\right] \tag{9}$$

Also, let us denote by $\boldsymbol{C}_{k,p}^{u}$ and by $\boldsymbol{C}_{k,p}^{l}$ the $NM \times (NM+1)$-dimensional matrix obtained by respectively retaining only the upper and lower part of the matrix $\boldsymbol{C}_{k,p}$, i.e.:

$$\boldsymbol{C}_{k,p}^{u} = [\boldsymbol{I}_{NM} \;\; \boldsymbol{O}_{NM}]\,\boldsymbol{C}_{k,p} \qquad \boldsymbol{C}_{k,p}^{l} = [\boldsymbol{O}_{NM} \;\; \boldsymbol{I}_{NM}]\,\boldsymbol{C}_{k,p} \tag{10}$$

Based on equations (6), (8) and (10), it is easy to show that the vector $\boldsymbol{r}_q(p)$ in (7) can be re-written as:

$$\boldsymbol{r}_q(p) = \sum_{k=0}^{K-1} \left(b_k(p-1)\boldsymbol{C}_{k,p-1}^{l} + b_k(p)\boldsymbol{C}_{k,p}^{u}\right)\boldsymbol{g}_{k,q} + \boldsymbol{n}_q(p) \qquad q = 0, \dots, Q-1 \tag{11}$$

with, in turn, $\boldsymbol{g}_{k,q}$ the following $(NM+1)$-dimensional vector:

$$\boldsymbol{g}_{k,q} = \left[\underbrace{0, \dots, 0}_{i_k}, \alpha_{k,q} A_k e^{j\phi_k}(1 - \delta_k M/T_c), \alpha_{k,q} A_k e^{j\phi_k} \delta_k M/T_c, \underbrace{0, \dots, 0}_{NM-i_k-1}\right] \tag{12}$$

Additionally, if we let

$$\overline{\boldsymbol{C}}_{k,p} = \left[\boldsymbol{C}_{k,p-1}^{l} \;\; \boldsymbol{C}_{k,p}^{u}\right] \qquad \overline{\boldsymbol{B}}_k(p) = \begin{bmatrix} b_k(p-1) \\ b_k(p) \end{bmatrix} \otimes \boldsymbol{I}_{NM+1} \tag{13}$$

the observable $\boldsymbol{r}_q(p)$ in (11) can be also expressed as:

$$\boldsymbol{r}_q(p) = \sum_{k=0}^{K-1} \overline{\boldsymbol{C}}_{k,p}\overline{\boldsymbol{B}}_k(p)\boldsymbol{g}_{k,q} + \boldsymbol{n}_q(p) \tag{14}$$

Interestingly, equation (11) reveals that, even though aperiodic codes are being adopted, the k-th user discretized signature in any signaling interval can be deemed as the product of a *known* time-varying matrix (i.e. $\overline{\boldsymbol{C}}_{k,p}\overline{\boldsymbol{B}}_k(p)$), times an unknown, *but* time-invariant, quantity (i.e. $\boldsymbol{g}_{k,q}$); additionally, from equations (11) and (12) it is also seen that all of the unknown parameters to be estimated have been lumped in the vectors $\boldsymbol{g}_{k,q}$'s, $\forall k = 0, \dots, K-1$ and $\forall q = 0, \dots, Q-1$. The algorithms that we are going to propose exploit the structure of such equations, and indeed they first estimate the vectors $\boldsymbol{g}_{k,q}$'s, and then furtherly process these estimates so as to come up with the estimates of the parameters $\boldsymbol{\theta}$, $\boldsymbol{\Phi}$, $\boldsymbol{a}$ and $\boldsymbol{\tau}$.

3.1 Decentralized (single-user) parameter estimation

Let us at first assume that we are interested in estimating the parameters from a given user only, the h-th say, i.e. we are interested in the estimation of the quantities τ_h, ϕ_h,

θ_h and A_h. For instance, such a situation may occurr in an asynchronous uplink CDMA channel, wherein a new user is starting its transmission and its parameters are to be acquired by the base station, or, alternatively, in a mobile station, which is obviously interested in acquiring the relevant parameters of the signal to be demodulated only.

The following assumptions are here made: a) the receiver has an initial uncertainty on the delay τ_h equal to T_b[1]; b) the bits $b_h(0), \dots, b_h(N_p)$, are known to the receiver, with N_p the number of signaling intervals devoted to the transmission of the known pilot symbols; c) no knowledge is assumed on the interfering signals, namely the number of interfering users, their signatures, delays, transmitted information bits, DoA's, phase offsets and amplitudes are not known to the receiver.

Based on the previous assumptions, we propose to resort to a Least-Squares-based procedure, i.e. the estimate, $\widehat{\boldsymbol{g}}_{h,q}(n)$ say, of the vector $\boldsymbol{g}_{h,q}$, available upon transmission of n known training symbols, is obtained as the solution to the following minimization problem:

$$\widehat{\boldsymbol{g}}_{h,q}(n) = \arg\min \sum_{p=1}^{n} \frac{1}{n} \left\| \boldsymbol{r}_q(p) - \overline{\boldsymbol{C}}_{h,p} \overline{\boldsymbol{B}}_h(p) \widehat{\boldsymbol{g}}_{h,q}(n) \right\|^2 \tag{15}$$

Solving problem (15) yields:

$$\widehat{\boldsymbol{g}}_{h,q}(n) = \left(\sum_{p=1}^{n} \frac{1}{n} \overline{\boldsymbol{B}}_h^H(p) \overline{\boldsymbol{C}}_{h,p}^H \overline{\boldsymbol{C}}_{h,p} \overline{\boldsymbol{B}}_h(p) \right)^{-1} \left(\sum_{p=1}^{n} \frac{1}{n} \overline{\boldsymbol{B}}_h^H(p) \overline{\boldsymbol{C}}_{h,p}^H \boldsymbol{r}_q(p) \right) \tag{16}$$

If we assume that the training sequence consists on N_p symbols, the estimate of the vectors $\boldsymbol{g}_{h,q}$ is obtained by evaluating expression (16) for $n = N_p$, $\forall q = 0, \dots, Q-1$. Since the product $\overline{\boldsymbol{B}}_h^H(p) \overline{\boldsymbol{C}}_{h,p}^H \overline{\boldsymbol{C}}_{h,p} \overline{\boldsymbol{B}}_h(p)$ is a square matrix of order $(NM+1)$, such a task entails a computational complexity equal to $\mathcal{O}\left((NM+1)^3\right)$; this might be an unaffordable computational effort, especially in mobile terminals, or, alternatively, if the transmitting station adopts a separate channel on which training symbols are continously transmitted, so that the estimate of the vectors $\boldsymbol{g}_{h,q}$ may be updated on a bit-interval basis. It is obviously understood that in this case a lower-complexity, possibly recursive, estimation procedure would be needed. To this end, we propose to adopt here a steepest-descent procedure, yielding the following recursive adaptation rule:

$$\widehat{\boldsymbol{g}}_{h,q}(n) = \left[\boldsymbol{I}_{NM+1} - \mu \sum_{p=1}^{n} \frac{1}{n} \overline{\boldsymbol{B}}_h^H(p) \overline{\boldsymbol{C}}_{h,p}^H \overline{\boldsymbol{C}}_{h,p} \overline{\boldsymbol{B}}_h(p) \right] \widehat{\boldsymbol{g}}_{h,q}(n-1) + \mu \sum_{p=1}^{n} \frac{1}{n} \overline{\boldsymbol{B}}_h^H(p) \overline{\boldsymbol{C}}_{h,p}^H \boldsymbol{r}_q(p) \tag{17}$$

with μ a suitably chosen real step-size. Since the product matrix $\overline{\boldsymbol{B}}_h^H(p) \overline{\boldsymbol{C}}_{h,p}^H \overline{\boldsymbol{C}}_{h,p} \overline{\boldsymbol{B}}_h(p)$ is highly structured, it can be easily shown that it can be computed with a quadratic computational complexity, whence the complexity per bit-interval of the rule (17) is $\mathcal{O}\,(NM)^2)$. Now, based on the estimated vectors $\widehat{\boldsymbol{g}}_{h,q}(n)$, we are to estimate the h-th user parameters A_h, τ_h, θ_h and ϕ_h. In principle, one could resort to a Maximum-Likelihood procedure, i.e. evaluate the likelihood of the vectors $\widehat{\boldsymbol{g}}_{h,q}(n)$ and maximize it with respect to the unknown parameters. Unfortunately, such a procedure is analitically unfeasible, unless the multiaccess interference is modeled as a colored Gaussian Random Process. Even so, however, the maximization procedure with respect to the h-th user unknown parameters leads to quite complicated equations. In order to keep down the system complexity, we resort here to the procedure below. Even though no claim of optimality can be made for this algorithm, it can be easily verified that it has the intuitively appealing property that, if the vectors $\boldsymbol{g}_{h,q}$ have been perfectly estimated $\forall q$, it yields perfect (i.e. with no error) estimates for the h-th user parameters.

---

[1] Observe that this is a very mild assumption, in that we are here requesting to have an initial estimate affected by an error contained in the interval $[-T_b/2, T_b/2]$, and such an estimate can be easily obtained by resorting, for instance, to a GPS-based common synchronization clock.

Step 1: Consider the following $(NM+1)$-dimensional vector:

$$\boldsymbol{g}_h = \sum_{q=0}^{Q-1} \left(\widehat{\boldsymbol{g}}_{h,q} \odot \widehat{\boldsymbol{g}}_{h,q}^* \right) \tag{18}$$

Denote by $(\boldsymbol{g}_h)_m$ the m-th entry of the vector $\boldsymbol{g}_h$, consider the NM-dimensional vector $\boldsymbol{g}'_h$ whose m-th entry is defined as:

$$(\boldsymbol{g}'_h)_m = (\boldsymbol{g}_h)_m + (\boldsymbol{g}_h)_{m+1} \qquad m = 1, \ldots, NM \tag{19}$$

and denote by $i'_h \in \{1, \ldots, NM\}$ the index of the maximum of the vector $\boldsymbol{g}'_h$. Let $\widehat{i}_h = i'_h - 1$.

Step 2: Obtain the estimate of the delay τ_h as:

$$\widehat{\tau_h} = \frac{\widehat{i}_h T_c}{M} + \frac{T_c}{MQ} \sum_{q=0}^{Q-1} \frac{\left(\widehat{\boldsymbol{g}}_{h,q}\right)_{\widehat{i}_h+1}}{\left(\widehat{\boldsymbol{g}}_{h,q}\right)_{\widehat{i}_h+1} + \left(\widehat{\boldsymbol{g}}_{h,q}\right)_{\widehat{i}_h}} \tag{20}$$

Step 3: Obtain an estimate of the h-th user' signal DoA through the equations:

$$\begin{gathered} \widehat{\alpha}_h = \frac{1}{Q-1} \sum_{q=0}^{Q-2} \frac{\left[\left(\widehat{\boldsymbol{g}}_{h,q+1}\right)_{\widehat{i}_h} \boldsymbol{c}_{h,p}^{(\widehat{i}_h)} + \left(\widehat{\boldsymbol{g}}_{h,q+1}\right)_{\widehat{i}_h+1} \boldsymbol{c}_{h,p}^{(\widehat{i}_h+1)} \right]^H \left[\left(\widehat{\boldsymbol{g}}_{h,q}\right)_{\widehat{i}_h} \boldsymbol{c}_{h,p}^{(\widehat{i}_h)} + \left(\widehat{\boldsymbol{g}}_{h,q}\right)_{\widehat{i}_h+1} \boldsymbol{c}_{h,p}^{(\widehat{i}_h+1)} \right]}{\left\| \left(\widehat{\boldsymbol{g}}_{h,q+1}\right)_{\widehat{i}_h} \boldsymbol{c}_{h,p}^{(\widehat{i}_h)} + \left(\widehat{\boldsymbol{g}}_{h,q+1}\right)_{\widehat{i}_h+1} \boldsymbol{c}_{h,p}^{(\widehat{i}_h+1)} \right\| \left\| \left(\widehat{\boldsymbol{g}}_{h,q}\right)_{\widehat{i}_h} \boldsymbol{c}_{h,p}^{(\widehat{i}_h)} + \left(\widehat{\boldsymbol{g}}_{h,q}\right)_{\widehat{i}_h+1} \boldsymbol{c}_{h,p}^{(\widehat{i}_h+1)} \right\|} \\ \widehat{\theta_h} = \arcsin\left(\frac{\lambda}{2\pi d} \arg(\widehat{\alpha}_h^*) \right) \end{gathered} \tag{21}$$

with $\arg(x)$ denoting the phase of the complex number x.

Step 4: Obtain an estimate of the complex gain $A_h e^{j\phi_h}$ as follows:

$$\begin{gathered} \widehat{\delta}_h = \widehat{\tau_h} - \frac{\widehat{i}_h T_c}{M} \\ \widehat{A_h e^{j\phi_h}} = \frac{1}{Q} \sum_{q=0}^{Q-1} (\widehat{\alpha}_h)^{-q} \frac{\left[(1-\widehat{\delta}_h) \boldsymbol{c}_{h,p}^{(\widehat{i}_h)} + \widehat{\delta}_h \boldsymbol{c}_{h,p}^{(\widehat{i}_h+1)} \right]^H \left[\left(\widehat{\boldsymbol{g}}_{h,q}\right)_{\widehat{i}_h} \boldsymbol{c}_{h,p}^{(\widehat{i}_h)} + \left(\widehat{\boldsymbol{g}}_{h,q}\right)_{\widehat{i}_h+1} \boldsymbol{c}_{h,p}^{(\widehat{i}_h+1)} \right]}{\left\| (1-\widehat{\delta}_h) \boldsymbol{c}_{h,p}^{(\widehat{i}_h)} + \widehat{\delta}_h \boldsymbol{c}_{h,p}^{(\widehat{i}_h+1)} \right\|^2} \end{gathered} \tag{22}$$

3.2 Centralized (Multiuser) parameter estimation

Let us now consider the case that the parameters of all of the active users are to be estimated, thus implying that all the users are transmitting known training symbols. Such a situation may occur in the uplink of an asynchronous CDMA-based packet netwowk, wherein users transmit a data packet in a given time-slot with a coarse synchronization, so that the signal received at the base station is actually asynchronous. Summing up, we are here assuming that: a) the receiver has an initial uncertainty on the delays $\boldsymbol{\tau}$ equal to $[-T_b/2, T_b/2]^K$; b) the bits $b_k(0), \dots, b_k(N_p)$, $\forall k = 0, \dots, K-1$ are known to the receiver, with N_p the number of known bits contained in the preamble of each packet.

Again, the estimation procedure that we are going to propose is a two-step one: first, based on a Least-Squares criterion, we estimate the vectors $\boldsymbol{g}_{k,q}$, $\forall k, q$, and then, we proceed to the estimation of the parameters $\boldsymbol{\tau}$, $\boldsymbol{\theta}$, $\boldsymbol{\Phi}$ and $\boldsymbol{a}$. To begin with, we observe that, if we let

$$\widetilde{\boldsymbol{C}}_p = \left[\overline{\boldsymbol{C}}_{0,p}\ \overline{\boldsymbol{C}}_{1,p}\ \dots\ \overline{\boldsymbol{C}}_{K-1,p}\right] \qquad \overline{\boldsymbol{B}}(p) = \mathrm{Diag}\left(\overline{\boldsymbol{B}}_0(p)\ \overline{\boldsymbol{B}}_1(p)\ \dots\ \overline{\boldsymbol{B}}_{K-1}(p)\right), \qquad (23)$$

and define $\widetilde{\boldsymbol{g}}_q = \left[\boldsymbol{g}_{0,q}^T\ \boldsymbol{g}_{1,q}^T\ \dots\ \boldsymbol{g}_{K-1,q}^T\right]^T$, the vector $\boldsymbol{r}_q(p)$ in (14) can be written as:

$$\boldsymbol{r}_q(p) = \widetilde{\boldsymbol{C}}_p \overline{\boldsymbol{B}}(p) \widetilde{\boldsymbol{g}}_q + \boldsymbol{n}_q(p) \qquad q = 0, \dots, Q-1 \qquad (24)$$

Accordingly, the estimate, $\widehat{\widetilde{\boldsymbol{g}}}_q(n)$ say, of the $K(NM+1)$-dimensional vector $\widetilde{\boldsymbol{g}}_q(n)$, available upon observation of the vectors $\boldsymbol{r}_q(1), \dots, \boldsymbol{r}_q(n)$, $\forall q$, solves the problem:

$$\widehat{\widetilde{\boldsymbol{g}}}_q(n) = \arg\min \sum_{p=1}^{n} \frac{1}{n} \left\| \boldsymbol{r}_q(p) - \widetilde{\boldsymbol{C}}_p \overline{\boldsymbol{B}}(p) \widehat{\widetilde{\boldsymbol{g}}}_q \right\|^2 \qquad (25)$$

which yields the solution:

$$\widehat{\widetilde{\boldsymbol{g}}}_q(n) = \left(\sum_{p=1}^{n} \frac{1}{n} \overline{\boldsymbol{B}}^H(p) \widetilde{\boldsymbol{C}}_p^H \widetilde{\boldsymbol{C}}_p \overline{\boldsymbol{B}}(p) \right)^{-1} \left(\sum_{p=1}^{n} \frac{1}{n} \overline{\boldsymbol{B}}^H(p) \widetilde{\boldsymbol{C}}_p^H \boldsymbol{r}_q(p) \right) \qquad (26)$$

It is worth pointing out that, for the case at hand, the least-squares criterion (25) is equivalent to the maximum-likelihood (ML) estimation procedure, whence equation (16) does represent the ML estimator of the vector $\widetilde{\boldsymbol{g}}_q$. Additionally, since it is seen from (24) that there exists a linear relationship between the noiseless data and the quantity to be estimated, it can be also shown that the equation (26) is also the Minimum Variance Unbiased Estimator (MVUE) for the vector $\widetilde{\boldsymbol{g}}_q$.

The computational complexity involved in the computation of expression (26) is $\mathcal{O}\left(K^3(NM+1)^3\right)$; also in this case alternative, lower-complexity, steepest-descent adaptation rules can be adopted, yielding:

$$\widehat{\widetilde{\boldsymbol{g}}}_q(n) = \left[\boldsymbol{I}_{K(NM+1)} - \mu \sum_{p=1}^{n} \frac{1}{n} \overline{\boldsymbol{B}}^H(p) \widetilde{\boldsymbol{C}}_p^H \widetilde{\boldsymbol{C}}_p \overline{\boldsymbol{B}}(p) \right] \widehat{\widetilde{\boldsymbol{g}}}_q(n-1) + \mu \sum_{p=1}^{n} \frac{1}{n} \overline{\boldsymbol{B}}^H(p) \widetilde{\boldsymbol{C}}_p^H \boldsymbol{r}_q(p) \qquad (27)$$

Again, since the matrix product $\overline{\boldsymbol{B}}^H(p) \widetilde{\boldsymbol{C}}_p^H \widetilde{\boldsymbol{C}}_p \overline{\boldsymbol{B}}(p)$ is a highly-structured square block-Hermitian matrix of order $K(NM+1)$, it can be easily shown that implementing the above adaptation rule entails a computational complexity equal to $\mathcal{O}\left((KNM)^2\right)$.

Now, the K vectors $\widehat{\boldsymbol{g}}_{k,q}(n)$ may be extracted from the $K(NM+1)$-dimensional vector $\widehat{\widetilde{\boldsymbol{g}}}_q(n)$, $\forall q = 0, \dots, Q-1$, and then the same steps as in the previous subsection can be followed to get estimates of the parameters τ_k, θ_k, ϕ_k and A_k, $\forall k = 0, \dots, K-1$.

4 Computer Simulations Examples

In this section we provide some simulation results illustrating the performance of the proposed adaptive algorithms. We have considered a CDMA System with processing gain $N = 15$, and with randomly generated spreading codes for all of the users. The number of samples per chip M has been set equal to 2. In figure 1, the performance of the proposed adaptive estimation algorithms has been reported versus the number of transmitted pilot symbols N_p. In particular, we assume that the user 0 is the one of interest and represent the following normalized square performance measures: a) Phase Tracking Square error: $\left(\frac{\widehat{\phi}_0 - \phi_0}{2\pi}\right)^2$; b) DoA Tracking Square error: $\left(\frac{\widehat{\theta}_0 - \theta_0}{2\pi}\right)^2$; c) Delay Tracking Square error: $\left(\frac{\widehat{\tau}_0 - \tau_0}{T_c}\right)^2$ and d) Amplitude Tracking Square error: $\left(\frac{\widehat{A}_0 - A_0}{A_0}\right)^2$. The solid line refers to the decentralized recursive estimation rule (17); the dashed line refers to the centralized recursive estimation rule (27), while, finally, the dot-dash line refers to the Cramer-Rao bound on the parameter estimation error (details on how to compute such a bound are not given due to lack of space). Additionally, we consider an array with $Q = 2$ antennas with spacing $d = \lambda/4$, there are $K = 6$ users, the user 0 received energy contrast $\gamma_0 = A_0^2 T_b / 2\mathcal{N}_0$ is equal to 14 dB and the Interference-to-Signal Ratio (ISR), i.e. the ratio A_h^2/A_0^2, $h = 1, \ldots, 5$, is set to 0 dB. Figure 2 considers the same system as figure 1, but with $Q = 4$. All of the represented plots are the result of an average over 400 independent runs with different realizations for the CDMA signals signatures and parameters. Inspecting the figures it is seen that the proposed method provides very satisfactory performance. In particular, as regards estimation of the angular parameters (i.e. phase offset and DoA) the centralized estimation procedure is very close to the CRB in both the considered scenario. As regards, instead, the estimation of the propagation delays and amplitudes, the steady-state square error is below 10^{-2}, i.e. the average accuracy is better than $T_c/10$ and $A_0/10$, respectively. As expected, the centralized estimation algorithm, which requires knowledge of spreading codes and transmitted bits by all of the active users, offers better performance than the decentralized procedure, even though such a performance gap is in some instances rather limited. Contrasting the results of figures 1 and 2 it is seen that increasing the antenna diversity order does not significantly improve the performance of the proposed estimation algorithms and the CRB's, with the only obvious exception of the DoA estimation error and CRB. Other simulation results, not shown here for the sake of brevity, have also revealed that the recursive estimation procedures (17) and (27), for suitably chosen values of the step size μ, exhibit a performance level very close to that of their non-recursive counterparts (16) and (26), respectively.

Summing up, the above results reveal that the newly derived algorithms are effective in estimating the received CDMA signals parameters. However, since in a communication link the ultimate performance measure is the Bit-Error-Rate (BER), it would be interesting to compare the BER of a multiuser detector built upon the parameter estimates furnished by our algorithm with that of the same multiuser detector built upon the true parameters values. Indeed, this would permit assessing, in terms of BER, the loss due to the prior uncertainty as to the received signals parameters. Accordingly, we have considered a system with no antenna diversity ($Q = 1$), $N = 15$, ISR= 0dB and $K = 6$. In figure 4 we have represented the user 0 system BER versus γ_0 for the classical linear one-shot minimum mean square error (MMSE) receiver [9], which requires knowledge of the users propagation delays, amplitudes and phase offsets, and for an adaptive MMSE receiver built upon the estimates of the said parameters. More precisely, at the beginning of each data packet, all of the users transmit N_p training bits; the centralized recursive estimation algorithm (27) is thus employed in order to come up with estimates of the CDMA signals parameters, which are subsequently employed in order to build the MMSE adaptive receiver. Interestingly, it is seen that, at 10^{-3} BER, the loss due to the prior uncertainty on the channel parameters is about 3dB for $N_p = 30$, while, for $N_p \geq 60$, the loss reduces to less than one dB. In figure 5, finally, we consider the same system as in figure 4, with the only exception that ISR= 6dB. As expected, in such a more severe scenario the adaptive receiver exhibit a poorer performance. Nonetheless, for $N_p \geq 60$, the performance loss with respect to the linear MMSE receiver is less than 2dB at a 10^{-3} system BER.

References

[1] J.G. Proakis, *Digital Communications*, 3rd edition, McGraw-Hill, 1996.

[2] E.G. Strom, S. Parkvall, S.L. Miller, B.E. Ottersten, "Propagation Delay Estimation in Asynchronous Direct-Sequence Code-Division Multiple Access Systems," *IEEE Trans. on Communications*, Vol. 44, pp. 84-93, January 1996.

[3] S.E. Bensley, B. Aazhang, "Maximum-Likelihood Synchronization of a Single User for Code-Division Multiple-Access Communication Systems," *IEEE Trans. on Communications*, Vol. 46, pp. 392-399, March 1998.

[4] U. Madhow, "Blind Adaptive Interference Suppression for the near-Far Resistant Acquisition and Demodulation of Direct-Sequence CDMA Signals," *IEEE Trans. on Signal Processing*, Vol. 45, pp. 102-112, January 1997.

[5] H.V. Poor, "On Parameter Estimation in DS/CDMA Formats," in *Advances in Communications and Signal Processing*, W.A. Porter and S.C. Kak Editors. New York: Springer-Verlag, 1989, pp.59-70.

[6] Z. Xie, C.K. Rushforth, R.T. Short, T.K. Moon, "Joint Signal Detection and Parameter Estimation in Multiuser Communications," *IEEE Trans. on Communications*, Vol. 41, pp. 1208-1215, August 1993.

[7] E. Dahlman, P. Beming, J. Knutsson, F. Ovesjö, M. Persson, C. Roobol, "WCDMA – The Radio Interface for Future Mobile Multimedia Communications," *IEEE Trans. on Vehicular Technology*, Vol. 47, pp. 1105-1118, November 1998.

[8] P. Taaghol, B. G. Evans, R. De Gaudenzi, G. Gallinaro, J. H. Lee, C. G. Kang, "Satellite UMTS/IMT2000 W-CDMA Air Interfaces," *IEEE Communications Magazine*, Vol. 37, pp.116-126, September 1999.

[9] S. Verdú, *Multiuser Detection*, Cambridge University Press: Cambridge (UK), 1998.

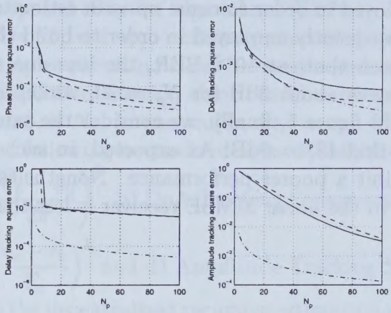


Figure 1: Performance of the centralized (dashed line) and decentralized (solid line) recursive estimation algorithms and CRB (dot-dashed line). System parameters: $\gamma_0 = 14$dB, $N = 15$, $Q = 2$, $K = 6$, ISR= 0dB.

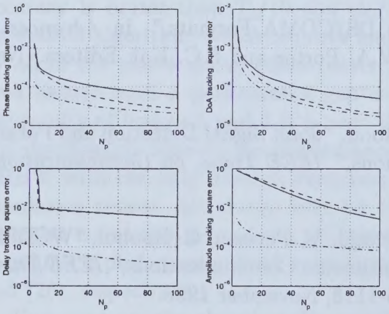


Figure 2: Performance of the centralized (dashed line) and decentralized (solid line) recursive estimation algorithms and CRB (dot-dashed line). System parameters: $\gamma_0 = 14$dB, $N = 15$, $Q = 4$, $K = 6$, ISR= 0dB.

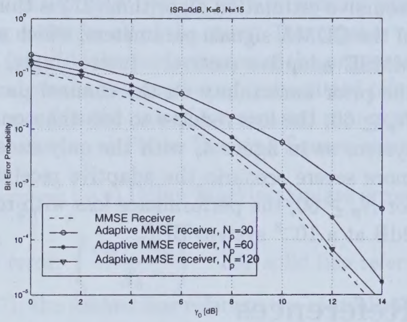


Figure 3: System BER for the linear MMSE receiver and for the adaptive MMSE detector. System parameters: $Q = 1$, $N = 15$, $K = 6$, ISR= 0dB.

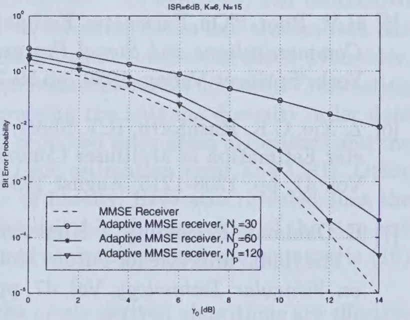


Figure 4: System BER for the linear MMSE receiver and for the adaptive MMSE detector. System parameters: $Q = 1$, $N = 15$, $K = 6$, ISR= 6dB.

Mobility Increases the Capacity of Ad-hoc Wireless Networks

Matthias Grossglauser
AT&T Labs- Research
180 Park Avenue
Florham Park NJ 07932
mgross@research.att.com

David Tse
Department of EECS
University of California
Berkeley CA 94720
dtse@eecs.berkeley.edu

Abstract

The capacity of ad-hoc wireless networks is constrained by the mutual interference of concurrent transmissions between nodes. We study a model of an ad-hoc network where n nodes communicate in random source-destination pairs. These nodes are assumed to be mobile. We examine the per-session throughput for applications with loose delay constraints, such that the topology changes over the time-scale of packet delivery. Under this assumption, the per-user throughput can increase dramatically when nodes are mobile rather than fixed. This improvement can be achieved by exploiting node mobility as a type of *multiuser diversity*.

1 Introduction

A fundamental characteristic of mobile wireless networks is the time variation of the channel strength of the underlying communication links. Such time variation can be due to multipath fading, path loss via distance attenuation, shadowing by obstacles and interference from other users. The impact of such time variation on the design of wireless networks permeates throughout the layers, ranging from coding and power control at the physical layer to cellular handoff and coverage planning at the networking layer.

An important means to cope with the time-variation of the channel is the use of *diversity*. Diversity can be obtained over time (interleaving of coded bits), frequency (combining of multipaths in CDMA systems) and space (multiple antennas). The basic idea is to improve performance by having several independent signal paths between the transmitter and the receiver.

These diversity modes pertain to a point-to-point link. Recent results point to another form of diversity, inherent in a wireless network with multiple users. This *multiuser diversity* is best motivated by an information theoretic result of Knopp and Humblet [4]. They focused on the uplink in the single cell, with multiple users communicating to the basestation via time-varying channels. To maximize the total information theoretic capacity, they showed that the optimal strategy is to schedule at any one time only the user with the best channel to transmit to the basestation. Diversity gain arises from the fact that in a system with many users, there is likely to be a user with a very good channel at any one time. Overall system throughput is maximized by allocating at any time the common channel resource to the user that can best exploit it.

Strategies of this type incur additional delay, because packets have to be buffered until the channel becomes strong relative to other users. Therefore, the time-scale of channel fluctuations that can be exploited through multiuser diversity is limited by the delay tolerance of the user or application. For example, for applications that can tolerate delays on the order of fractions of seconds to several seconds, short time-scale multipath fading can be taken advantage of. In this paper, the focus is on applications that are so asynchronous in nature that they can tolerate end-to-end delays of minutes or even hours. On such a long time-scale, even more diversity gain can be obtained because the *network topology* changes significantly over time due to user mobility. Examples of such applications include electronic mail, database synchronization between a mobile terminal and a central database, and certain types of event notification.

We demonstrate in this paper that these ideas have ramifications to the design of wireless networks beyond classical cellular architectures. We will focus on mobile ad-hoc networks which have no fixed basestations and with multiple pairs of users wanting to communicate with each other. Gupta and Kumar [2] proposed a model for studying the capacity of *fixed* ad-hoc networks, where nodes are randomly located but are immobile. Each source node has a random destination to which it wants to communicate. Their main result shows that as the number of nodes per unit area n increases, the throughput per source-destination (S-D) pair decreases approximately like $1/\sqrt{n}$. This is the best performance achievable even allowing for optimal scheduling, routing and relaying of packets in the networks, and is a somewhat pessimistic result on the scalability of such networks, as the traffic rate per S-D pair actually goes to zero.

In this paper, we introduce mobility into the model and consider the situation when users move independently around the network. Our main result shows that the average long-term throughput per S-D pair can be kept *constant* even as the number of nodes per unit area n increases. This is in sharp contrast to the fixed network scenario, and the dramatic performance improvement is obtained through the exploitation of the time-variation of the users' channels due to mobility. We observe that our result implies that, at least in terms of growth rate as a function of n, there is no significant loss in throughput per S-D pair when there are many nodes in the network as compared to having just a single S-D pair. A caveat of this result is that the attained long-term throughput is averaged over the time-scale of node mobility, and hence delays of that order will be incurred.

In the fixed ad-hoc network model, the fundamental performance limitation comes from the fact that long-range direct communication between many user pairs is infeasible, due to the excessive interference caused. As a result, most communication has to occur between nearest neighbors, at distances of order $1/\sqrt{n}$, with each packet going through many other nodes (serving as relays) before reaching the destination. The number of hops in a typical route is of order $\sqrt{n}$. Because much of the traffic carried by the nodes are relayed traffic, the actual useful throughput per user pair has to be small.

With mobility, a seemingly natural strategy to overcome the above performance limitation is to transmit only when the source and destination nodes are close together, at distances of order $1/\sqrt{n}$. This is reminiscent of the Infostation architecture [1], where users connect to the infostations only when they are close by. However, this strategy turns out to be too naive in the present situation. The problem is that the fraction of time two nodes are nearest neighbors is too small, of the order of $1/n$. Instead, our strategy is for each source node to distribute its packets to as many different nodes as possible. These then serve as mobile relay nodes and whenever they get close to the final destination, they hand the packets off to the final destination. The basic idea is that since there are many different relay nodes, the probability that at least one is close to the destination is significant. On the other hand, each packet goes through at most one relay node, and hence the throughput can be kept high. Although the basic communication problem is point-to-point, this strategy effectively creates multiuser diversity by distributing packets to many different intermediate nodes which have independent time-varying channels to the final destination.

2 Model

The ad-hoc network consists of n nodes all lying in the open disk of unit area (of radius $1/\sqrt{\pi}$). The location of the ith user at time t is given by $X_i(t)$. Nodes are mobile, and we assume that the process $\{X_i(\cdot)\}$ is stationary and ergodic with stationary distribution uniform on the open disk; moreover the trajectories of different users are independent and identically distributed.

We first describe the session model. We assume that each of the n nodes is a *source* node for one session, and a *destination* node for another session. Let us stipulate that the source node i has data intended for destination node $d(i)$. We assume that each source node has an infinite reservoir of packets to send to its destination. The source-destination association does not change with time, although the nodes themselves move.

We next describe the transmission model. At (slotted) time t, let $P_i(t)$ be the transmit power of node i, and $\gamma_{ij}(t)$ be the channel gain from node i to node j, such that the received power at node j is $P_i(t)\gamma_{ij}$. At time t, node i transmits data at rate R packets/sec to node j if

$$\frac{P_i(t)\gamma_{ij}(t)}{N_0 + \frac{1}{L}\sum_{k\neq i} P_k(t)\gamma_{ij}(t)} > \beta, \tag{1}$$

where β is the signal-to-interference ratio (SIR) requirement for successful communication, N_0 is the background noise power, and L is the *processing gain* of the system. For a narrowband system, $L = 1$, while for a spread-spectrum CDMA system, L is larger than 1. In this paper, we only consider large-scale path loss characteristics in the fading channel model. The channel gain is given by

$$\gamma_{ij}(t) := \frac{1}{|X_i(t) - X_j(t)|^\alpha},$$

where α is a parameter greater than 2.

We will consider two models. In the first model, packets are only transmitted directly from the source to the destination, without relaying through other nodes. In other words, $j(i) = d(i)$. In the second model, nodes can serve as *relays* for packets intended for other nodes. We assume that each node has an infinite buffer to store such packets.

At any time t, a *scheduler* chooses which nodes will be senders, and the power levels $P_i(t)$ for these senders. The objective of the scheduler for both models is to ensure a high long-term throughput for each S-D pair. More precisely, consider a scheduling and relay policy π. Let $M_i^\pi(t)$ be the number of source node i packets that destination $d(i)$ receives at time t under policy π. Given the random trajectories of the users, we shall say a long term throughput of $\lambda(n)$ is feasible if there is a policy π such that for every S-D pair i,

$$\liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} M_i^\pi(t) \geq \lambda(n).$$

We note that the throughput $\lambda(n)$ is a random quantity as it depends on the random locations of the users. The indexing by the system size n emphasizes that we are interested in studying the asymptotic behavior as n becomes large.

3 Results

3.1 Fixed Nodes

First we review results of Gupta and Kumar [2]. They focus on a static model, where nodes are not mobile. The node positions $\{X_i\}$ are i.i.d. and uniformly distributed in the open disk of unit area, but fixed over time. The destination for each source node is chosen to be the node closest to a randomly chosen point on the disk. The destinations are all chosen independently. Relaying is allowed in their model. The following results yield upper and lower bounds on the asymptotically feasible throughput.

Theorem 3.1 *(Main result 4 in [2]) There exists constants c and c′ such that*

$$\lim_{n\to\infty} Pr\left\{ \lambda(n) = \frac{cR}{\sqrt{n\log n}} \quad \textit{is feasible} \right\} = 1,$$

and

$$\lim_{n\to\infty} Pr\left\{ \lambda(n) = \frac{c'R}{\sqrt{n}} \quad \textit{is feasible} \right\} = 0,$$

Thus, within a factor of $\sqrt{\log n}$, the throughput per S-D pair goes to zero like $R/\sqrt{n}$ in the case when the nodes are fixed.

3.2 Mobile Nodes Without Relaying

The reason why the throughput for fixed nodes goes to zero is that the number of relay nodes a packet has to go through scales as $\sqrt{n}$. However, in our model of mobile nodes, any two nodes can be expected to be close to each other every from time to time. This suggests that we may be able to improve the capacity by not relaying at all, and only letting sources transmit directly to destinations. We now show that without relaying, there is no way to achieve an $0(1)$ throughput per S-D pair.

We first need the following Lemma. This fact is already established in the proof of Theorem 2.1(ii) in [2], but we include the proof here for completeness.

Lemma 3.2 *Consider a scheduling policy that schedules direct transmissions only. Fix an arbitrary time t. Let $\mathcal{S}(t)$ be the set of source nodes that are scheduled successful transmission to their respective destinations. Then*

$$\sum_{i\in\mathcal{S}(t)} |X_i(t) - X_{j(i)}(t)|^\alpha \le B,$$

where

$$B := 2^\alpha \pi^{-\alpha/2} \frac{\beta + L}{\beta}.$$

This Lemma shows that the number of simultaneous long-range communication is limited by interference. Since the distance between the source and destination is $O(1)$ most of the time, this limitation in turn puts a bound on the performance of any strategy which uses only direct communication.

Theorem 3.3 *Assuming that the policy is only allowed to schedule direct transmission between the source and destination nodes, and no relaying is permitted. If c is any constant satisfying*

$$c > \left[2^{\alpha}\left(1+\frac{2}{\alpha}\right)\pi^{-\alpha/2}\frac{\beta+L}{\beta}\right]^{\frac{1}{1+\alpha/2}},$$

then

$$Pr\left\{\lambda(n) = cn^{-\frac{1}{1+\alpha/2}}R \quad \textit{is feasible}\right\} = 0$$

for sufficiently large n.

This result says that without relaying, the achievable throughput per S-D pair goes to zero at least as fast as $n^{-\frac{1}{1+\alpha/2}}$.

3.3 Mobile Nodes With Relaying

The problem of only using direct transmission between the source and destination is that too much of the communication is long-range, and the resulting interference limits the number of concurrent transmissions. This suggests that the total throughput of the network can be increased if we constrain transmission to nearest neighbors. Indeed, Theorem 3.4 below demonstrates that $O(n)$ *concurrent successful transmissions* per time slot are possible. The problem with this approach is that the source and the destination node of a session are nearest neighbors only for a very small fraction of time, on the order of $1/n$, vanishing as the user density increases. Therefore, although we can schedule $O(n)$ communicating sender-receiver pairs per time slot, the fraction of pairs that actually has a packet to transmit vanishes for large n.

The idea now is to spread out packets to a large number of intermediate *relay nodes* that temporarily buffer packets until final delivery to the destination is possible. For a source-destination pair S-D, all the other $n-2$ nodes can serve as relay nodes. The goal is that in steady-state, the packets of every source node will be distributed across all the nodes in the network, hence ensuring that every node in the network will have packets buffered destined to every other node (except itself). This ensures that a scheduled sender-receiver pair always has a packet to send, in contrast to the case of direct transmission.

The question is how many times a packet has to be relayed in order to spread traffic uniformly to all nodes. In fact, as the node location processes $\{X_i(t)\}$ are independent, stationary and ergodic, it is actually sufficient to *relay only once*. This is because the probability for an arbitrary node to be scheduled to receive a packet from a source node S is equal for all nodes and independent of S. Each packet then makes two hops, one from the source to its random relay node, and one from that relay node to the destination. As no packet is transmitted more than twice, the achievable total throughput is $O(n)$.

We now make the above argument rigorous. We first exhibit a scheduling policy π to select random sender-receiver pairs in each time slot t, such that all the pairs can successfully transmit in time slot t. We will then use this policy as a building block to achieve $O(1)$ throughput per S-D pair for large n.

The scheduling policy π is as follows. Let us focus on a particular time slot t. To simplify notation, we will drop the time index t in the following discussion. We fix a *sender density* parameter $\theta \in (0, 1)$. We randomly designate $n_S = \theta n$ of the nodes as senders in each time slot, and the remaining n_R nodes as *potential* receivers. Specifically, we randomly pick one out of $\begin{pmatrix} n_S \\ n \end{pmatrix}$ equally likely partitions of the n nodes into the set of senders $\mathcal{S}$ and the set of potential receivers $\mathcal{R}$. Each sender node transmits packets to its nearest neighbor *among all nodes in* $\mathcal{R}$, using unit transmit power ($P_i = 1$). Among the n_S sender-receiver pairs, we retain those for which the interference generated by the other senders is sufficiently small that transmission is possible. Let N_t be the number of such pairs. Theorem 3.4 below shows that the number of feasible sender-receiver pairs N_t is $O(n)$. Note that the set of sender-receiver pairs is random and that it depends only on the node locations $\{X_i\}$.

Theorem 3.4 *For the scheduling policy* π, *the expected number* $E[N_t]$ *of feasible sender-receiver pairs is* $O(n)$, *i.e.,*

$$\lim_{n\to\infty} \frac{E[N_t]}{n} = \phi > 0. \tag{2}$$

Furthermore, for two arbitrary nodes i *and* j, *the probability that* (i, j) *is scheduled as a sender-receiver pair is* $O(1/n)$.

We can now apply this scheduling policy π to our basic problem. The overall algorithm is divided into two phases: (1) scheduling of packet transmissions from sources to relays (or the final destination), and (2) scheduling of packet transmissions from relays (or the source) to final destinations. These two phases are interleaved: in the even time-slots, phase 1 is run; in the odd time-slots, phase 2 is run.

In phase 1, we can apply the scheduling policy π to transmit packets from sources to relays or destinations. In phase 2, we again apply the policy π, but this time to transmit packets from relays to final destinations (or, as in phase 1, from a source directly to the destination). More specifically, when a receiver is identified for a sender under π, the sender checks if it has any packets for which the receiver is the destination; if so, it will transmit it. It should be noted that every packet goes through at most two hops: it is transmitted once in phase 1 from its source to an intermediate relay, and once in phase 2 from a relay to the final destination. We allow for packets to be directly transmitted from their source to their destinations in both phases, if a sender-receiver pair happens to be a source-destination pair as well.

Let us analyze the throughput per S-D pair under this two-phased scheme. As π only depends on node locations and because the node locations $\{X_i(t)\}$ are i.i.d., stationary, and ergodic, the long-term throughput between any two nodes is equal to the probability that these two nodes are selected by π as a feasible sender-receiver pair. According to Theorem 3.4, this probability is $O(1/n)$. Now, for a given S-D pair, there is one direct route and $n - 2$ two-hop routes which go through one relay node R. The throughput over the direct route is $O(1/n)$. For each two-hop route, we can consider the relay node R as a single server queue (cf. Fig. 1). Applying Theorem 3.4, we see that both the arrival rate and the service rate of this queue is the same and $O(1/n)$. Summing over the throughputs of all the $n-1$ routes, it can be seen that the total average throughput per S-D pair is $O(1)$. We have proved the following Theorem, which is the main result of this paper.

Theorem 3.5 *The two-phased algorithm achieves a throughput per S-D pair of* $O(1)$, *i.e. there exists a constant* $c > 0$ *such that*

$$\lim_{n\to\infty} Pr\{\lambda(n) = cR \quad is\, feasible\} = 1.$$

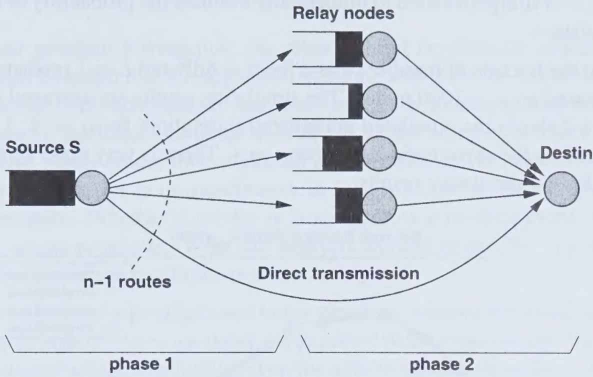


Figure 1: The two-phase scheduling policy viewed as a queueing system, for a source-destination pair: in phase 1, a packet at S is served by a queue of capacity $O(1)$, and is forwarded either to the destination or to one of $n-2$ relay nodes with equal probability. The service rate at each relay node R is $O(1/n)$, for a total session rate of $O(1)$.

Note that the largest possible throughput is $c = \phi/2$.

The essence of the proof of Theorem 3.4, and the fundamental reason why we can have $O(n)$ concurrent nearest neighbor transmission, is the fact that the received power at the nearest neighbor is of the same order as the total interference from $O(n)$ number of interferers. A similar phenomenon has been observed in [3] in the cellular setting, where they have shown that, provided $\alpha > 2$, the capture probability of the nearest transmitter to the basestation does not go to zero as the number of interferers become large. Although these results may seem surprising on first sight, they are all based on the following property: if $W_1, \ldots, W_n$ are i.i.d. random variables such that the cdf $F(w)$ decays slower than w^{-1} as $w \to \infty$, then the largest of them is of the same order as the sum. In the context of our problem, W_i's are the received powers from the transmitting nodes.

The technical complication in the proof of Theorem 3.4 is due to the fact that both the distribution of the received power from the sender and the distribution of the interference depends on the location of the receiver. This is primarily due to the edge effects of the disk, and this dependency would not be present if for example the nodes are randomly located on the surface of a sphere. Fortunately, in the regime we are interested in, the asymptotic distributions depends only on what happens in the local neighborhood around the receiver, and this is independent of where the receiver is in the open disk.

4 Numerical Results

We have examined the throughput capacity both through numerical evaluation of the asymptotic probability of capture developed in the preceding section, and through simulation of random network topologies.

We have evaluated the asymptotic fraction of feasible pairs ϕ for the special case $\alpha = 4$, because for this case, it is straightforward to numerically evaluate the probability of capture through Monte-Carlo simulation.

We have compared the fraction of feasible pairs ϕ for $\beta = 6$dB and $L = 1$ predicted by our model with simulations based on $n = 1000$ nodes. The simulation results are averaged over 20 random topologies. Figure 2 shows the simulated normalized throughput for $\alpha = 2$, 3, and 4, and the throughput predicted by the asymptotic model for $\alpha = 4$. There is very good agreement between the analytical model and simulation results.

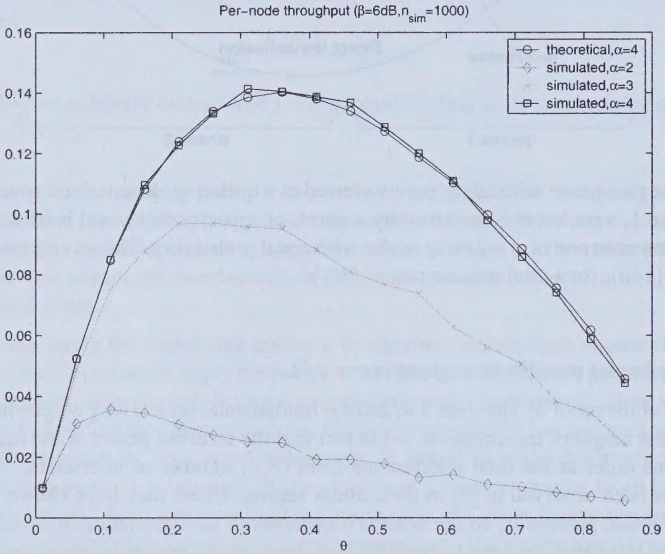


Figure 2: The normalized per-node throughput, as a function of the sender density θ, for different values of α. For $\alpha = 4$, the throughput predicted by the model is also shown.

It is evident from the figure that given α, there exists an optimal sender density θ that maximizes the throughput. If θ is too small, then we do not exploit the potential for spatial channel reuse. If θ is too large, then the interference power becomes too dominant. The optimal θ obviously depends on α. For small α, interference limits the spatial channel reuse. Hence, the sender density has to be small. For large α, interference is more localized, and the optimal θ and the maximum throughput are larger.

5 Discussion

5.1 Distributed Implementation

Although in our problem formulation, we allow central coordinated scheduling, relaying and routing, it should be noted that the algorithm obtained above can be implemented in completely distributed manner. At each time instant, each node can randomly and independently decide whether it wants to participate in phase 1 or phase 2, and whether it wants to be a sender or a potential receiver. Each sender then seeks out a potential receiver nearest to it, and attempts to send data to it. The access is uncoordinated; in fact, multiple senders may attempt to transmit to the same receiver. Whether a sender is "captured" is a random event, much like standard MAC random access protocols. What our analysis showed is that the probability of success is reasonable even in a network with many users.

Note that the two-phased algorithm used in the proof was chosen for mathematical convenience. As the capacity in both phases is identical, the expected delay experienced by a packet from source to destination would actually be infinite even for a finite number of nodes n. It is straightforward to fix this problem, e.g., by allowing both S-R and R-D transmissions to occur concurrently, but giving absolute priority to R-D (phase 2) transmission in all scheduled sender-receiver pairs. A detailed study of local scheduling strategies and their impact on end-to-end delay is the subject of future work.

5.2 Related Work

Frenkiel *et al.*, in [1], have introduced the concept of an *Infostation* for delay-insensitive data applications. An Infostation is a high-speed wireless basestation that does away with ubiquitous coverage. The motivation is that if delay is unimportant, then capacity is maximized by using the entire transmit power budget when the user is close to the basestation, and no power when the user is far away. While the work on Infostations has motivated us to look at delay-insensitive applications, our focus is different. We are not chiefly concerned about maximizing the capacity of a point-to-point channel subject to a power budget, but instead we are interested in the degree of spatial reuse in an entire network of mobile nodes that all desire to communicate, and we propose a new type of multiuser diversity to achieve high overall capacity.

Hajek, Krishna and LaMaire have studied a related interference model [3]. They examine the probability of capture for a single receiver for an asymptotically large number of senders. They show that in the limit, this probability only depends on the roll-off exponent α, but not on other channel effects such as fading and shadowing. While their results are not directly applicable to our setting, we nevertheless believe that this robustness property to other channel effects carries over.

In their numerical experiments, they report that for $\alpha = 4$ and $\beta = 6$dB, the asymptotic probability of capture is approximately 0.319. In our setting, as can be seen from Fig. 2, the probability of capture is approximately 0.14. The reason for this difference is that we study a sender-centric scheduling policy, while Hajek *et al.* study the probability of capture of a single receiver.

Finally, as discussed earlier, Gupta and Kumar [2] have studied the capacity of random, but fixed ad-hoc networks. Their main insights are that communication must be limited to "near" neighbors in order to permit dense spatial channel reuse. However, the throughput per session goes to zero as the number of nodes n scales up, because each session requires on the order of $\sqrt{n}$ hops. A similar problem setup has also been considered by Shepard [5].

6 Conclusion

In this paper, we have examined the asymptotic throughput capacity of wireless ad-hoc networks. We have identified a new type of multiuser diversity that exploits long time-scale fluctuations in the fading process due to node mobility. Our results show that direct communication between sources and destinations is not sufficient to exploit this diversity, because they are too far apart most of the time. We propose to spread traffic randomly to intermediate relay nodes to take advantage of additional "routes" between a source and a destination. A single relay node is sufficient to use the entire throughput capacity of the network within the limits imposed by the interference model. This explains the dramatic performance improvement over a fixed ad-hoc network, where $O(\sqrt{n})$ intermediate relay nodes are necessary.

The key assumption in our model is that an end-to-end packet delivery delay on the order of the *mobility time-scale* must be acceptable. This is clearly not the case for voice communication. Hence, wireless telephony cannot benefit from the multiuser diversity inherent in node mobility. However, wireless data services are expected to grow quickly over the next few years. A subset of these services, such as email and database synchronization, do indeed possess very loose delay constraints (on the order of hours). Also, wireless devices are bound to become smaller and more pervasive in the future; they will not only be carried by humans, but integrated into physical objects (such as cars, electrical appliances, etc.) It is unlikely that the density of base-stations will keep pace, due to regulatory and environmental hurdles in deploying them. Thus, there is a clear opportunity for wireless ad-hoc networks to extend the reach of wireless communication. Our results suggest that delay-tolerant applications can take advantage of node mobility to significantly increase the throughput capacity of such networks.

References

[1] Richard Frenkiel, B. R. Badrinath, Joan Borras, and Roy Yates. The Infostations challenge: Balancing cost and ubiquity in delivering wireless data. *Preprint*, August 1999.

[2] Piyush Gupta and P. R. Kumar. The Capacity of Wireless Networks. *IEEE Transactions on Information Theory*, 46(2):388–404, March 2000.

[3] B. Hajek, A. Krishna, and R.O. LaMaire. On the capture probability for a large number of stations. *IEEE Transactions on Communications*, 45(2):254–260, February 1997.

[4] R. Knopp and P. A. Humblet. Information Capacity and power Control in Single-Cell Multiuser Communications. In *Proc. Int. Conf. on Communications*, Seattle, June 1995.

[5] Timothy J. Shepard. A Channel Access Scheme for Large Dense Packet Radio Networks. In *Proc. ACM SIGCOMM '96*, San Francisco, Calif., August 1996.

Some Capacity Results for Ad Hoc Networks *

S. Toumpis and A. Goldsmith

Stanford University

Abstract

We present results on the capacity of ad hoc networks in two different settings. In the first, all nodes wish to communicate with a single common node, possibly using multiple hops, power control and time division. Bounds on the performance of the network are derived in the form of a capacity region, and it is shown that power control only marginally improves the performance. In the second setting, we develop a similar framework for the case when nodes wish to communicate with arbitrary collections of other nodes.

1 Introduction

Wireless networks consist of a number of nodes communicating over a wireless channel. Depending on their architecture, they can be roughly divided in two categories. In those following the *cellular* paradigm (Fig. 1(a)), all nodes communicate directly with a base station that is responsible for controlling all transmissions and forwarding packets to the intended users. In those following the *ad hoc* paradigm (Fig. 1(b)) all nodes have the same capabilities and responsibilities. Two nodes wishing to communicate do so directly or use nodes lying in between them to route their packets. Our work deals with this second type of networks.

Ad hoc networks first attracted significant research interest in the early '80s [5]. At that time technology constraints greatly limited the performance of ad hoc networks. However, recent advances in the technology of semiconductors, signal processing,modulation and coding, and networking protocols, along with an increased interest in wireless applications, has led to a renewed interest in ad hoc networking.

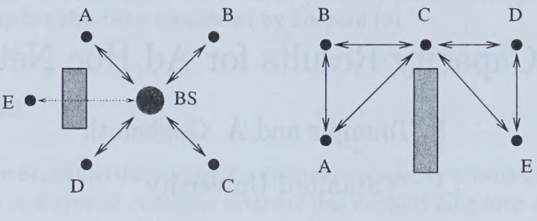
The nature of the wireless channel, the lack of synchronization, and also the lack of any predetermined topology creates many challenging research topics in this area. On the network layer, routing is now a much more challenging issue, since the topology changes dynamically, and also because the wireless channel is interference limited [6].

On the medium access control (MAC) layer, the lack of centralized control implies that random access protocols such as CSMA and flavors of ALOHA are more favorable. However, these suffer from the well known problems of "hidden" and "exposed" terminals. The authors of [4] and [1] propose a "4 way handshake" algorithm that deals with the issue. This algorithm was adopted in the design of the IEEE802.11 wireless LAN protocol [2].

Lately, there has also been work on determining the capacity of ad hoc networks [3]. The authors of [3] derived upper bounds on the performance of a class of networks in

---

*Work supported by the Office of Naval Research (ONR), grant number N00014-99-1-0698.



(a) (b)

Figure 1: Examples of the two paradigms of wireless networks. In the cellular network of (a), nodes A-D exchange information by routing through the base station. Node E cannot communicate with any other node, because of an obstacle that blocks the line of sight to the base station. In the ad hoc network of (b) any two nodes can communicate if a path between them exists. Nodes A and E will have to use node C as a relay for their communication.

the limit of a large number of nodes. Here we introduce a framework for studying the capacity of ad hoc networks with any number of nodes.

The rest of the paper is organized as follows: In Section 2 we present our assumptions regarding the nodes and the channel. In Section 3 we define and study the capacity region of a network in which all nodes wish to relay information to a common node, possibly by using multiple hops. In section 4 we extend the framework of Section 3 to treat the general case, in which nodes wish to communicate with one or more arbitrarily chosen nodes. Throughout the paper, terms being defined are set in **boldface**.

2 Assumptions

Let $\mathcal{A} = \{A_1, A_2, \dots, A_n\}$ be a collection of n nodes. Each node A_i is equipped with a receiver, a transmitter and an infinite buffer, and wishes to communicate with some of the other nodes, possibly by routing information packets through intermediate nodes, and/or using time division. We assume that nodes do not wish to broadcast information, so every transmitted packet is intended for a single node. We also assume that nodes cannot transmit and receive at the same time.

Node A_i transmits with power P_i and all transmissions occupy the full bandwidth W of the system. We define the **power vector** to be the vector $P = [P_1 \; P_2 \; \dots \; P_n]^T$. When node A_i transmits with power P_i, node A_j receives the signal with power $G_{ij}P_i$. We define the **channel gain matrix** to be the following $n \times n$ matrix:

$$G = \begin{bmatrix} 0 & G_{12} & \dots & G_{1n} \\ G_{21} & 0 & \dots & G_{2n} \\ \dots & \dots & \dots & \dots \\ G_{n1} & G_{n2} & \dots & 0 \end{bmatrix} \tag{1}$$

The receiver of each node is subject to thermal noise and interference from other users. We assume the noise is stationary with power η_i for node A_i and define the **noise vector** $H = [\eta_1 \; \eta_2 \; \dots \; \eta_n]^T$.

Let $\{A_t : t \in \mathcal{T}\}$ be the set of transmitting nodes at a given time. We assume that node A_t transmits with power P_t. Let us assume that node A_j is receiving information from node $A_i, i \in \mathcal{T}$. Then the **signal to interference and noise ratio (SINR)** at node A_j will be

$$\gamma_{ij} = \frac{G_{ij}P_i}{\eta_j + \sum_{k \in T, k \neq i} G_{kj}P_j}. \tag{2}$$

We assume that the signal received by A_j is of acceptable quality if A_i sets the transmission rate R_{ij} to satisfy

$$R_{ij} \leq f(\gamma_{ij}) \tag{3}$$

where $f(.)$ is a function characteristic of the receiver and the performance metric. For example, under a Shannon analysis we set

$$f(\gamma_{ij}) = W \log_2(1 + \gamma_{ij}). \tag{4}$$

Then the communication is reliable if the rate is set to be less than the Shannon capacity of the link. Alternatively, $f(\gamma_{ij})$ could correspond to the maximum data rate to meet a given BER requirement, under a specific modulation scheme. Since transmitting with rate smaller than the maximum is suboptimal, we assume that all nodes transmit with the rate defined in Eqn. 3 with equality

Finally, we assume that the nodes have perfect knowledge of the channel gain matrix (G) and the noise (H) and power (P) vectors, and have agreed on the channel access method.

3 Capacity Regions of Centralized Ad Hoc Networks

3.1 Centralized ad hoc networks

In this section we concentrate on networks in which all nodes wish to forward information to a common node. Specifically, let $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$ be a collection of n nodes for which the assumptions of Section 2 hold and additionally $A_1, \ldots, A_{n-1}$ wish to forward information only to A_n. Then we say the collection is a **centralized ad hoc network**. We concentrate on this network not only because its analysis is tractable but also because it is an accurate model of a number of physical networks, for example cellular networks that allow the users to forward their packets to the base station via other users.

Example 1 *In the centralized ad hoc network of Fig. 2, nodes A_1, A_2 and A_3 wish to send information to A_4. Since there is no line of sight between A_1 and A_3, and between A_2 and A_4, it is reasonable to set:*

$$G = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \tag{5}$$

We also set $P = [10\ 10\ 10\ 10]^T$, $H = [1\ 1\ 1\ 1]^T$ and $W = 10\{\log_2(\frac{21}{11})\}^{-1}$. and assume that the maximum transmission rate is limited by the Shannon capacity limit of Eqn. 4.

3.2 Definitions

Before discussing the capacity regions of centralized networks, we need to develop a succint way of describing their behavior. Let us assume that at time t a total of k nodes are transmitting. We denote these by $\{A_{T(1)}, \ldots, A_{T(k)}\}$ (so that the $T(i)$'s are

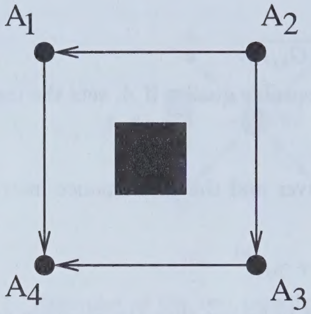
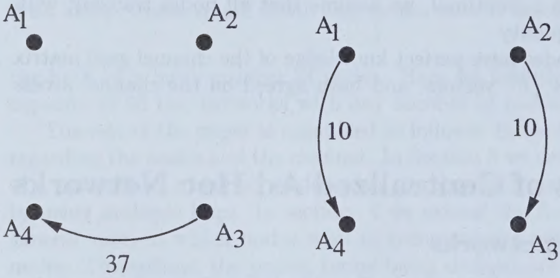


Figure 2: The centralized ad hoc network of Example 1. Because of the obstacle, nodes A_1 and A_3, and nodes A_2 and A_4 cannot communicate directly.



(a) (b)

Figure 3: Two transmission modes of the centralized ad hoc network of Example 1: (a)M_1. (b)M_2.

distinct integers). Each of these nodes will have an intended receiver. Let $A_{R(i)}$ be the intended receiver of $A_{T(i)}$ (so that the $R(i)$'s are also distinct integers, all different from the $T(i)$'s). We will call **links** the k vectors $l_i = (A_{T(i)}, A_{R(i)})$, $i = 1, \ldots, k$. So the links are k pairs containing each of the transmitters and their respective receiver. We also define the **transmission mode** of the network at time t to be the set of all active links $M = \{l_1, \ldots, l_k\}$.

Continuing Example 1, the setting of Fig. 3.2(a) (where A_3 transmits to A_4) can be succintly described by the transmission mode $M_1 = \{(A_3, A_4)\}$ and the setting of Fig. 3.2(b) (where A_1 transmits to A_4 and A_2 transmits to A_3) by the transmission mode $M_2 = \{(A_1, A_4), (A_2, A_3)\}$. The rates of communication are also shown on the figures. The rates are calculated by using the Shannon capacity bound of Eqn. 4 and the parameters P, H, W given in the example.

Transmission modes are complete descriptions of the state of a centralized network at a given time t. However, as time progresses the group of transmitters and/or their intended receivers will change. In other words, the transmission mode of the network will change with time, and a complete description of the evolution of a network can be given in terms of the modes that are used and the time intervals during which these modes are active. We can limit the discussion to networks that alternate periodically between a given set of modes. Each transmission mode is used for a specified time, and once all modes are used, the schedule is repeated.

Formally, let $\mathcal{A} = \{A_1, \ldots, A_n\}$ be a centralized ad hoc network and $M_1, \ldots, M_k$ a

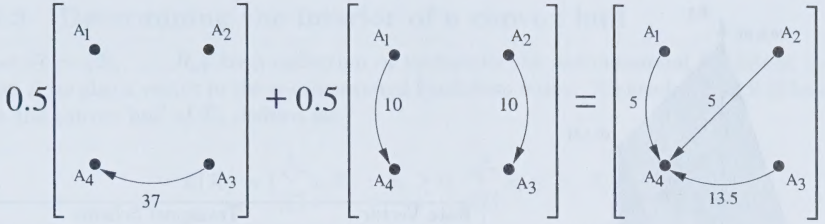


Figure 4: Schedule $\mathcal{S}$ as a linear combination of modes M_1 and M_2: $\mathcal{S} = 0.5M_1 + 0.5M_2$

number of transmission modes. We will say that the network will operate under the **transmission schedule**

$$\mathcal{S} = \sum_{i=1}^{k} \tau_i M_i \ , \ \sum_{i=1}^{k} \tau_i = 1 \tag{6}$$

if the network alternates periodically between the modes M_i and remains in them for a time percentage specified by the τ_i's.

In the setting of Example 1, we will say that the network operates under the transmission schedule $\mathcal{S} = 0.5M_1 + 0.5M_2$ if half of the time A_3 transmits to A_4 and half of the time A_1 transmits to A_4 and A_2 transmits to A_3.

Transmission modes and schedules are two constructions that allow us to describe the behavior of networks as time progresses. However, to discuss the capacity region of networks, we need to convert these constructions into tangible numbers. Therefore, we will introduce the notion of **rate vectors**. To give the motivation behind the definition, we study the network of Example 1. We see that, under schedule $\mathcal{S}$, A_1 transmits $5\frac{bits}{sec}$ to A_4, A_2 transmits $5\frac{bits}{sec}$ to A_4, and A_3 transmits $13.5\frac{bits}{sec}$ to A_4. Therefore, the rate vector of the scheme is [5 5 13.5]. Note that the rate vector can be written as a weighted sum of the rate vectors of the modes, weighted by their time percentages:

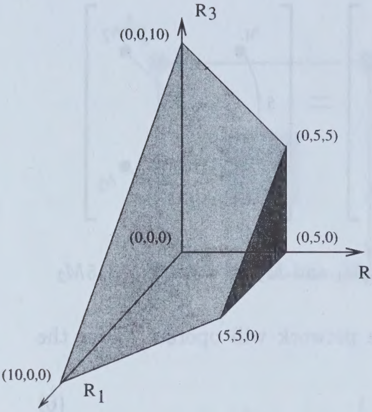
$$[5\ 5\ 13.5] = 0.5[0\ 0\ 37] + 0.5[10\ 10\ -10] \tag{7}$$

The vectors [0 0 37] and [10 10 −10] correspond to the modes M_1 and M_2 and denote that rates with which nodes A_1, A_2 and A_3 send information. Negative rate values correspond to receiving packets from other users, as we explain in more detail below.

We now introduce rate vectors formally. Let $\mathcal{A} = \{A_1, \ldots, A_n\}$ be a centralized ad hoc network and M a transmission mode. If, under this mode, node A_i forwards information to any node with rate $r(i)$, then the **rate vector R_M of transmission mode** M is defined as the $(n-1)$-dimensional vector $[r(1)\ r(2) \ldots r(n-1)]$. A negative entry implies that the respective node actually *receives* information with rate $-r(i)$. Also, let $M_1, M_2, \ldots, M_k$ be a number of transmission modes, with their rate vectors being $R_{M_1}, R_{M_2}, \ldots, R_{M_k}$, and a transmission schedule $\mathcal{S} = \sum_{i=1}^{k} \tau_i M_i$. The **rate vector of the transmission schedule** $\mathcal{S}$ is defined to be $R_{\mathcal{S}} = \sum_{i=1}^{k} \tau_i M_i$.

We note that the rate vectors of some modes, such as that of mode M_2 of Example 1 contain negative entries. Those entries imply that, when the network operates under these modes, the respective nodes receive, information from other nodes that subsequently must be forwarded. When a network operates continuously under these modes, these nodes will accumulate information that was really intended for A_n. Such modes must be combined with other modes to form transmission schemes whose rate vectors have no negative entries.

The rate vector of a transmission schedule represents a combination of rates with which nodes $A_1, \ldots, A_{n-1}$ can communicate with A_n using this schedule. Therefore, the



Rate Vector	Transport Scheme
[0 0 0]	$\emptyset$
[10 0 0]	$\{(A_1, A_4)\}$
[0 0 10]	$\{(A_3, A_4)\}$
[0 5 0]	$0.5\{(A_2, 3)\} + 0.5\{(A_3, A_4)\}$
[5 5 0]	$0.5\{(A_1, A_4), (A_2, A_3)\} + 0.5\{(A_3, A_4)\}$
[0 5 5]	$0.5\{(A_3, A_4), (A_2, A_1)\} + 0.5\{(A_1, A_4)\}$

(a) (b)

Figure 5: (a)The capacity region of the centralized network of Example 1. (b) The vertices of the capacity region of Example 1 and the schemes that achieve them.

collection of all possible rate vectors corresponding to a transmission scheme forms the capacity region of the network. Since rate vectors of transmission schedules with negative components must be excluded (as was discussed in the previous paragraph), we formally define the **capacity region** of a centralized ad hoc network with nodes $A_1, \ldots, A_n$ to be the set

$$C(\mathcal{A}) = \{R_{\mathcal{S}}\} \cap \mathcal{P}_{n-1} \tag{8}$$

where $\mathcal{P}_{n-1}$ is the set of all vectors of size $(n-1)$ with non-negative entries.

To further characterize the capacity region of a centralized network, we note that the number of all distinct combinations of receivers and transmitters is finite. Therefore the number of different transmission modes is finite. For example, in the case of centralized ad hoc networks of size 3 there are only 5 different modes, including the one in which all nodes remain silent. Let $\{M_1, M_2, \ldots, M_K\}$ be an enumeration of all the modes of the centralized network $\mathcal{A}$. Since the rate vector of *any* scheme may be written as the linear combination of rate vectors corresponding to a set of modes, with the coefficients being positive and summing to unity, an alternative description for the capacity region is

$$C(\mathcal{A}) = \text{Convex Hull}\{R_{M_i} : 1 \leq i \leq K\} \cap \mathcal{P}_{n-1} \tag{9}$$

Since K is finite, the capacity region will generally be a convex polyhedron in the $(n-1)$-dimensional space that can be concisely described in terms of its vertices. In Fig. 5(a) we plot the capacity region of the network of Example 1. In Fig. 5(b) we list the vertices of the capacity region and the transmission schedules that achieve them.

Eqn. 9 implies that we can determine whether a rate vector belongs to the capacity region by solving the most basic problem in computational geometry, specifically the problem of whether a vector belongs to the convex hull of a set of vectors. In the next paragraph we show one way of solving this problem, which additionally supplies us with the linear combination of vertices that achieve the rate vector in question.

3.3 Determining the interior of a convex hull

Let $\mathcal{R} = \{R_1, \ldots, R_k\}$ be a collection of vectors in the n-dimensional Euclidean space. Let R be also a vector in the n-dimensional Euclidean space. We need to find if R belongs to the convex hull of $\mathcal{R}$, defined as:

$$\mathcal{C}(\mathcal{R}) = \{\sum_{i=1}^{k} x_i R_i \ : \ x_i \geq 0, \ \sum_{i=1}^{k} x_i \leq 1, \ R_i \in \mathcal{R}\} \tag{10}$$

Let $A = [R_1 \ R_2 \ \ldots \ R_k]$. A will be a matrix of size $n \times k$. Let $I_{k\times k}$ be the unit matrix of size $k \times k$, $O_{k\times 1}$ the matrix of size $k \times 1$ and zero elements, and $F_{1\times k}$ the matrix of size $1 \times k$ and unit elements. Let x_M be a matrix of size $1 \times k$ that is the solution of the following linear programming problem:

$$\begin{aligned} x_M &= \text{argmin}_x\{F_{k\times 1}x\} \\ \text{subject to:} \quad &- I_{k\times k}x_M \leq O_{k\times 1}, \ Ax_M = R \end{aligned} \tag{11}$$

A necessary and sufficient condition for R to belong to the convex hull, is that a solution to (11) exists, and $\sum_{i=1}^{k} x_M(i) \leq 1$. The linear programming problem can be solved by use of standard routines, such as those in MATLAB.

3.4 Effects of power control, multiple hops and simultaneous transmissions

We considered the impact of the following features on the capacity region of centralized ad hoc networks:

1. Allowing the nodes to adapt their transmitter power.
2. Allowing the nodes to forward the packets of other nodes.
3. Allowing many nodes to transmit at the same time.

In our study, we create 20 independent centralized ad hoc networks of size 7 by placing nodes $A_1, \ldots, A_6$ randomly, uniformly and independently around the centralized node A_7. The channel gain matrix is obtained by assuming that the channel gains between any two nodes decay exponentially with distance, and also depend on a lognormal component.

Since the capacity regions of these networks are polyhedra in the 6-dimensional space and are hard to describe, we use as a figure of merit for the efficiency of the network the maximum uniform rate with which all nodes can communicate with the central node.

We first consider networks in which no power control, routing or simultaneous transmissions are allowed, and choose a value for W such that the uniform rate is around 1. Because of the randomness in the channel gain matrix, the rate fluctuates as much as 50%.

We then allow nodes to use routing. This significantly increases the uniform rate, since now single hop routes with high attenuation are dropped in favor of multihop routes will low attenuation. The uniform rates in this case are around 10, but fluctuate as much as 30% between different networks.

As a next step, we also allow simultaneous transmissions for different links, in addition to routing. Under these assumptions the uniform rates increase to around 13, with a 30% fluctuation. The intuition behind this result is that whenever a node transmits, there is another node that can also transmit, that will not strongly interfere with the first transmission.

Finally, in addition to routing and simultaneous transmissions we allow nodes to use power control. Specifically, we allow nodes to use k different levels of transmitter power, and increase k until the maximum uniform rate converges. We find that power control, when used in addition to routing and simultaneous transmissions, only marginally improves the performance of centralized networks, by about 5% at most. An interpretation of this result is that using time division is a type of generic power control that is very close to optimal.

The above trends hold for networks of various sizes. Furthermore, allowing simultaneous transmissions has a more dramatic effects on the performance of larger networks.

4 Capacity Regions of Ad Hoc Networks

The formulation of the previous section can be generalized in the case of **ad hoc networks** that follow the assumptions of Section 2 and in which every node wishes to communicate with an arbitrary collection of the rest of the nodes. To reflect this generalization, we will have to use slightly different definitions for links, transmission modes, transmission schedules, and the capacity region.

Let $\mathcal{A} = \{A_1, A_2, \ldots, A_n\}$ be a collection of nodes forming an ad hoc network. We concentrate on the state of the network at time t. At this time, a set of nodes, which we index by $A_{T(i)}, i = 1, \ldots, k$, is transmitting. We assume that node $A_{T(i)}$ transmits information that was created by node $A_{S(i)}$ and is intended for node $A_{R(i)}$ (if node $A_{T(i)}$ is the original information source, then $A_{T(i)} = A_{S(i)}$).

We define as **links** the k vectors $l_i = (A_{S(i)}, A_{T(i)}, A_{R(i)})$, and will say that, at time t, the network operates under the **transmission mode** $M = \{l_1, \ldots, l_k\}$. In addition, we will say that the network will operate under the **transmission schedule** $\mathcal{S} = \sum_{i=1}^{k} \tau_i M_i$ if the network periodically alternates between the transmission modes M_i and remains in them for a time percentage specified by the τ_i's.

Example 2 *As in Example 1, we will consider the setup of Fig. 2. However in this setting each node is interested in transmitting information to an arbitrary set of nodes. We define the following two modes:*

$$M_1 = \{(A_1, A_2, A_1), (A_3, A_4, A_3)\} \qquad M_2 = \{(A_2, A_3, 1), (A_4, A_1, A_3)\} \tag{12}$$

When the network operates under mode M_1, A_1 sends its own information to A_2, and at the same time A_3 sends its own information to A_4. When the network operates under mode M_2, A_2 forwards A_1's information to A_3 and A_4 sends A_3's information to A_1. These modes are depicted in Figs. 6(a) and 6(b) respectively. The rates are calculated as in Example 1. In Fig. 6 we specify for each link the original source of the data transmitted, as well as the link rate. We can also define the transmission schedule $\mathcal{S} = 0.75M_1 + 0.25M_2$, according to which the network operates under mode M_1 for 75% of the time and under mode M_2 for the rest of the time. The schedule appears in Fig. 7.

To proceed with a definition of the capacity region of ad hoc networks, we need to introduce the idea of **rate matrices**, which are analogous to the rate vectors of the previous section. To motivate the definition, we note that under the transmission schedule of the previous example, A_1 transmits information with rate 5 to A_2 and with rate 2.5 to A_3, and A_3 transmits information with rate 5 to A_4 and rate 2.5 to A_1. All these rates appear in the following matrix:

$$R_{\mathcal{T}} = \begin{bmatrix} -7.5 & 5 & 2.5 & 0 \\ 0 & 0 & 0 & 0 \\ 2.5 & 0 & -7.5 & 5 \\ 0 & 0 & 0 & 0 \end{bmatrix} \tag{13}$$

$$= 0.75 \begin{bmatrix} -10 & 10 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -10 & 10 \\ 0 & 0 & 0 & 0 \end{bmatrix} + 0.25 \begin{bmatrix} 0 & -10 & 10 & 0 \\ 0 & 0 & 0 & 0 \\ 10 & 0 & 0 & -10 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad (14)$$

$$= 0.75 R_{M_1} + 0.25 R_{M_2} \quad (15)$$

In the above equation, $R_{\mathcal{T}}$ corresponds to the transmission schedule $\mathcal{S}$ and R_{M_1} and R_{M_2} correspond to the modes M_1 and M_2 respectively. In each of the above matrices, an entry with indices ij indicates the rate with which node A_j stores information that originated in node A_i.

Formally, let $\mathcal{A} = \{A_1, \ldots, A_n\}$ be a collection of nodes forming an ad hoc network. Let M be a transmission mode of this network, such that, when the network operates under this mode, node A_i gains information whose original source is A_j, with rate r_{ij}. Then the **rate matrix of the transmission mode** M is the $n \times n$ matrix defined as follows:

$$R_M = \begin{bmatrix} r_{11} & r_{12} & \ldots & r_{1n} \\ r_{21} & r_{22} & \ldots & r_{2n} \\ \ldots & \ldots & \ldots & \ldots \\ r_{n1} & r_{n2} & \ldots & r_{nn} \end{bmatrix} \quad (16)$$

If an entry r_{ij} is negative, under this mode node A_j is transmitting information that has originated in A_i with rate $-r_{ij}$

Let $\{M_1, \ldots, M_k\}$ be a collection of transmission modes. Let $R_{M_1}, \ldots, R_{M_k}$ be their rate matrices. Then the **rate matrix of the scheme** $\mathcal{S} = \sum_{i=1}^{k} \tau_i M_i$ is defined as the matrix $\sum_{i=1}^{k} \tau_i R_{M_i}$. The **capacity region** of the ad hoc network is defined to be the set of matrices

$$C = \{R_S\} \cap \mathcal{Q}_n \quad (17)$$

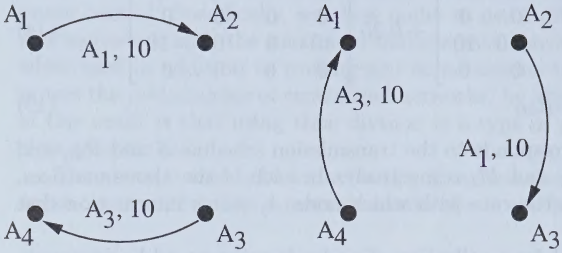
where $\mathcal{Q}_n$ is the set of $n \times n$ matrices whose elements off the main diagonal are non-negative. Since the space of $n \times n$ matrices is isomorphic to the n^2-dimensional Euclidean space, and since the sum of the elements along any row of any rate matrix is zero, the capacity region may be thought of as a set in the $(n^2 - n)$-dimensional Euclidean space. As in the previous section, we note that the number of distinct transmission modes is finite. If we index all different modes by $M_1, M_2, \ldots, M_K$, we derive the following alternative expression for the capacity region:

$$C(\mathcal{A}) = \text{Convex Hull}\{R_{M_i} : 1 \le i \le K\} \cap \mathcal{Q}_n \quad (18)$$

Similarly to the previous case, the capacity region is a convex polyhedron in the $n^2 - n$ dimensional space. Unfortunately, the high dimensionality of the problem does not allow a simple numerical analysis.

5 Summary

We develop a framework for studying the performance of ad hoc networks in two different settings. In the first setting all nodes wish to communicate with a single common node. In the second, each node wishes to communicate with an arbitrary collection of other nodes. In both cases, the transport capabilities of the networks are quantified in terms of the capacity regions, which are shown to be convex regions in the Euclidean space of dimension n and $(n^2 - n)$ respectively. In the first setting, it is shown that allowing many active transmissions at a given time as well as routing greatly enhances the performance of the network. However, the effects of power control are only marginal. Our current research effort is towards the establishment of similar results in the second setting.



(a) (b)

Figure 6: Two transmission modes of the centralized ad hoc network of Example 1: (a)M_1. (b)M_2.

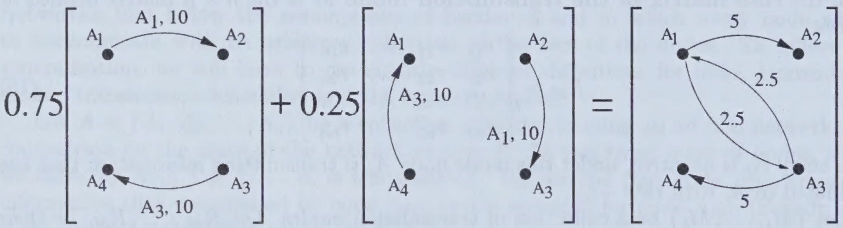


Figure 7: Schedule $\mathcal{S}$ as a linear combination of modes M_1 and M_2: $\mathcal{S} = 0.75M_1 + 0.25M_2$

References

[1] V. Bharghavan, A. Demers, S. Shenkar, and L. Zhang. Macaw: A media access protocol for wireless lan. *Proceedings of the SIGCOMM '94 Conference of Communications Architectures, Protocols and Applications*, pages 212–225, August 1994.

[2] IEEE Computer Society LAN MAN Standards Committee. *Wirelss LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications: IEEE Standard 802.11*. The Institute of Electrical and Electronics Engineers, New York, NY, 1997.

[3] Piyush Gupta and P. R. Kumar. The capacity of wireless networks. *IEEE Transactions on Information Theory*, 46(2), March 2000.

[4] P. Karn. Maca: A new channel access method for packet radio. *Proceedings of the 9th Computer Networking Conference*, pages 134–140, September 1990.

[5] B. Leiner, D. Nielson, and F. Tobagi. Issues in packet radio design. *Proceedings of the IEEE*, January 1987.

[6] Elizabeth M. Royer and C.-K. Toh. A review of current routing protocols for ad-hoc mobile wireless networks. *IEEE Personal Communications Magazine*, pages 46–55, April 1999.

Cooperative Detection and Communication in Wireless Sensor Networks

Mohiuddin Ahmed, Yung-Szu Tu, and Gregory Pottie
Electrical Engineering Department, University of California, Los Angeles
Box 951594, Los Angeles, California 90095, USA

Abstract— We define information theoretic problems in the cooperative detection of targets by a distributed wireless sensor network, and in the cooperative transmission of the results to a remote user. The dominant constraint in both cases is the energy cost of communications, rather than bandwidth or signal processing costs. We show that the detection problem can be cast as a rate distortion problem, and bound the region for Gaussian sources. We describe how the cooperative transmission problem differs from the usual context for space-time codes, and present one practical approach.

I. INTRODUCTION

In wireless sensor networks, potentially a very large number of devices cooperate to detect an event, and report the results of the detection exercise to some end user, who may be remote from the network. A fundamental constraint in sensor networks is the energy of the nodes, which may be limited in their lifetime energy reserves by batteries, and which likely have low peak power available for transmission. Questions then naturally arise as to the most efficient means to make reliable detection decisions, and in achieving longer range communications. In section II we set up the cooperative detection (data fusion) problem as a rate distortion problem, and in section III present the analytic formulation. In section IV we introduce the cooperative communication problem, and present some simple approaches. In section V we present capacity results, and in section VI suggest one simple coding approach that can apply to a large network with varying numbers of cooperating elements. In section VII we present our conclusions.

II. THE DATA FUSION PROBLEM

We now consider the question: what data rates can be supported for data fusion in a sensor network, given a specified tolerable data distortion? Unfortunately, an exact answer to this question is not available in the general case, and thus we have considered the specific case of correlated Gaussian sources in this paper.

The motivation for this problem comes from performing distributed detection of phenomena. It is well known from the theory of distributed detection that higher reliability and lower probability of detection error can be achieved when observation data from multiple, distributed sources is intelligently fused in a decision making algorithm, rather than using a single observation data set [1]. This, coupled with the fact that fabrication technological advances have made low-cost sensors incorporating wireless transceivers, signal processing and sensing in one integrated package a desirable low-cost option, it is inevitable that such devices will be widely used in detection applications such as security, monitoring, diagnostic, remote exploration etc. This has given rise to the development of wireless integrated networked sensors (WINS) [2], Figure 1.

However, the effective deployment of such distributed processing systems introduces some significant design issues, most notably: networking and communication protocols, transmission channel and power constraints, and scalability, among others [2], [3]. It is also evident that some fundamental limits are required to assess the optimality of any system design with regard to the "best design". Thus, an information theoretic analysis of the system is required. Here we assume that the primary constraint is power.

A WINS system invokes a multi-terminal analysis, as diagrammed in Figure 1.

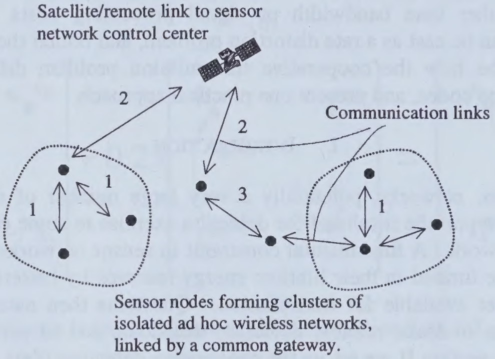

Satellite/remote link to sensor network control center
2
2
Communication links
1
1
3
1
Sensor nodes forming clusters of isolated ad hoc wireless networks, linked by a common gateway.

Figure 1: Wireless Integrated Network Sensor System

For this type of a system, all the traditional types of multi-terminal channels considered in information theory appear: the multiple access channel (communication pathways shown in Figure 1 numbered as 1), the broadcast channel (2), the relay and interference channel (3), etc. [4]. Additionally, the channel may be fading or more complex. Unfortunately, in the absence of a general information theory of multi-terminal networks, there is, as yet, no analytical way of evaluating performance bounds for whole systems, for a specific type of task for which the network might be employed, e.g. distributed detection. Our goal is to apply the results known so far to obtain, if not global optimum information limits, at least optimality criteria for each of the individual sub-blocks. In this regard, advances have been made with certain simplifying considerations, most notably the rate-distortion bounds for multiple, correlated nodes. The focus so far has been on pathways 1, inside the local loop.

For the individual local network loops, the problem is one of efficient communication and data fusion for detection. Associated problems such as network boot-strap, algorithms determining the minimum number of nodes necessary for reliable detection of a phenomenon, etc. have been studied and are not discussed here [3], [5]. Instead the coding problem is considered. The multi-terminal coding theory problem for two correlated memoryless sources with separate encoders has been solved by Slepian and Wolf [5]. The correlated sources assumption is valid in the WINS case, since for nodes observing the same target, the data generated for each sensor is expected to be correlated. Also, in the WINS case, it is apparent that power efficiency can be incorporated by allowing a distortion criteria, since there may be several data fusion centers, and since local processing provides a far higher power gain than RF transmissions. Thus, what becomes of interest then is how much distortion can be tolerated if the sensor network is to achieve some measure of efficiency in distributed detection – in other words, the rate-distortion bound.

Previous work in this area by Wyner and Ziv [7], Han and Kobayashi [9], [10] and Csizar [11] have all focused on special extensions of Slepian and Wolf, but the general rate-distortion regions characterization problem has remained unsolved. We have extended the special case for two correlated memoryless Gaussian sources (Oohama, [7]) to the n- sources case (with partial side information). In what follows, the notation is adopted from Oohama, [7]. The next section presents the analytic formulation of the problem and the main result.

III. ANALYTIC FORMULATION

Consider the multi-sensor system as shown below (Figure 2).

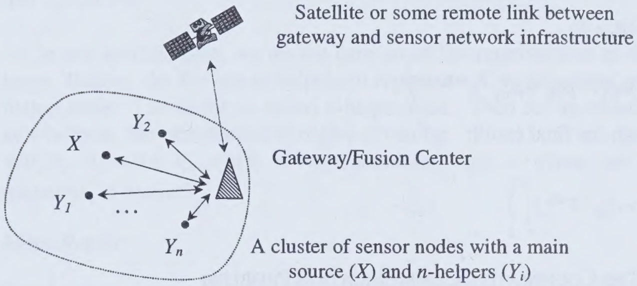

Satellite or some remote link between gateway and sensor network infrastructure
Y_2
X
Gateway/Fusion Center
Y_1
...
Y_n
A cluster of sensor nodes with a main source (X) and n-helpers (Y_i)

Figure 2: Data Fusion for a Wireless Networked Sensor System

A portion of a distributed cluster of sensor nodes (perhaps mobile) is observing a phenomenon and generating source data. Algorithms exist which can determine which nodes in the proximity of the phenomenon need to be activated and which can remain dormant [4]. Once this boot-up process is completed, the node observation data is assumed to be Gaussian (for analytical simplicity), with one data node acting as the main data source (e.g. that which is closest to the phenomenon), and the remaining nodes generating correlated data. The coding challenge is then to determine appropriate codes and data rates such that the gateway/data-fusion center can reproduce the data from the main node using the remaining nodes as sources of partial side information, subject to some distortion criteria.

Thus for a main source, X, and n correlated sources, Y_i, with $\{X_t, Y_{1t}, \cdots Y_{nt}\}_{t=1}^{\infty}$ being stationary Gaussian memoryless sources, for each observation time, t=1, 2, 3, ..., we let the random $(n+1)$-tuplet $(X_t, Y_{1t}, \cdots Y_{nt})$ take values in $X \times Y_1 \times \cdots Y_n$. The joint probability density function is given by the usual expression for the multi-dimensional Gaussian probability density function, where the covariance matrix can be denoted as:

$$\Lambda = \begin{bmatrix} \sigma_X^2 & \rho_{XY_1}\sigma_X\sigma_{Y_1} & \cdots & \rho_{XY_n}\sigma_X\sigma_{Y_n} \\ \rho_{XY_1}\sigma_X\sigma_{Y_1} & \sigma_{Y_1}^2 & \cdots & \rho_{Y_1Y_n}\sigma_{Y_1}\sigma_{Y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{XY_n}\sigma_X\sigma_{Y_n} & \rho_{Y_1Y_n}\sigma_X\sigma_{Y_1} & \cdots & \sigma_{Y_n}^2 \end{bmatrix} \quad (1)$$

We can write independent copies of $\{X_t\}_{t=1}^{\infty}$ as $X^m = X_1, X_2, \cdots, X_m$ and similarly for $\{Y_{kt}\}_{t=1}^{\infty}$, k=$1,2,\ldots,n$. Next, we consider a coding system where data sequences $X^m, \cdots Y_k^m$ are

$F_n(D) = \inf\limits_{\hat{X}^n : \Delta_1 \le D} \frac{1}{n} I\left(X^n; \hat{X}^n\right)$, $G_i(R) = \sup\limits_{\substack{W_i : \frac{1}{n} I\left(Y_i^n; W_i\right) \le R \\ W_i \to Y_i^n \to X^n}} \frac{1}{n} I\left(X^n; W_i\right)$ we get:

$$R_X + \delta \ge F_n(D_X + \delta) - \tfrac{1}{n} \sum_{k=1}^{n} G_k(R_k + \delta)$$

By the Gaussian property of the sources, and concavity of the logarithm function, we get: $\frac{1}{n} I\left(X^n; \hat{X}^n\right) \ge \frac{1}{2} \log\left(\frac{\sigma_X^2}{D}\right) \;\Rightarrow\; F_n(D) \ge \frac{1}{2} \log\left(\frac{\sigma_X^2}{D}\right)$

Finally, an upper bound of $G_i(R)$ is obtained by an extension of the technique for the two node case (entropy of the power inequality, monotonicity, Jensen's inequality): $\frac{1}{n} I\left(X^n; W_i\right) \le \frac{1}{2} \log\left(\frac{1}{1 - \rho_{XY_i}^2 + \rho_{XY_i}^2 \cdot 2^{-2R}}\right)$

Substituting the bounds in the expression for $F_n(D)$, we obtain:

$$\begin{aligned} R_X + \delta &\ge F_n(D_X + \delta) - \tfrac{1}{n} \sum_{k=1}^{n} G_k(R_k + \delta) \\ &\ge \frac{1}{2} \log\left(\frac{\sigma_X^2}{D_1 + \delta}\right) + \frac{1}{2} \sum_{k=1}^{n} \log\left(1 - \rho_{XY_k}^2 + \rho_{XY_k}^2 \cdot 2^{-2(R_k + \delta)}\right)^{\frac{1}{n}} \end{aligned}$$

Letting $\delta \to 0$, we obtain the final result:

$$R_X \ge \frac{1}{2} \log\left\{\frac{\sigma_X^2}{D_X} \cdot \left[\prod_{k=1}^{n} \left(1 - \rho_{XY_k}^2 + \rho_{XY_k}^2 \cdot 2^{-2R_k}\right)\right]^{\frac{1}{n}}\right\}.$$

IV. THE COOPERATIVE COMMUNICATION PROBLEM

With many low-power and low-cost sensors available, the goal is to achieve more robust and higher rate communications. For example, we may again consider the situation in Figure 1, where now communication to the satellite may be accomplished by a cooperating group of nodes to overcome their peak power limitations. This problem is of interest for example in remote planetary exploration, where the cost of providing power generation on the surface of a planet is orders of magnitude greater than for the downlink. The communication schemes to accomplish this goal must be insensitive to the number of transmitters and their location, and robust to the transceiver failure or motion. We are also interested in problems where for example clusters of nodes on the surface must cooperate to overcome gaps in the multi-hop network caused by infelicitous node placement, or node failures. Thus multiple transmitters and receivers may cooperate.

The simplest way to achieve cooperative communication is by multiplexing. Frequency, time, code division multiple access (FDMA, TDMA, CDMA), and multiuser orthogonal frequency division multiplex (multiuser OFDM) are all feasible approaches. Using these methods, channels starting from different transmitters can work in parallel and independently.

Different transmitters could cooperate with each other or process information independently. For FDMA, TDMA, if they process information independently, obviously, the improvement on the performance over the single transmitter system is that information data rate increases by a factor of the number of transmitters. Meanwhile, the bandwidth expands by the same factor. Even if coding is used across transmitters, the data rate increases by the same factor because the received symbol rate increases by the same factor. Although the analyses for CDMA and multiuser OFDM are not so straightforward, there is also an

separately encoded to $\varphi_1(X^m), \varphi_1(Y_k^m)$ and sent to the information processing / data fusion center. The decoder function, $\Psi = (\psi_X, \psi_1, \cdots, \psi_n)$, observes the *(n+1)*-tuplet $\{\varphi_1(X^m), \varphi_1(Y_1^m), \cdots, \varphi_1(Y_n^m)\}$ and estimates $\left(\hat{X}^n, \hat{Y}_1^m, \cdots, \hat{Y}_n^m\right)$. We let $\Im_{n,\delta}(R_X, R_1, \cdots, R_n)$ denote the set of all such coding and decoding schemes, $(\varphi_X, \varphi_1, \cdots, \varphi_n, \Psi)$, which can exist with the properties mentioned above. We take :
$d_x : X^2 \to [0,\infty),\ \ d_1 : Y_1^2 \to [0,\infty), \cdots, d_n : Y_n^2 \to [0,\infty)$
as the distortion measures, which, in our case, is the squared distortion measure, and we let the average distortions be $\Delta_1 = E\left\{\frac{1}{m}\sum_{t=1}^{m} d_1(X_t, \hat{X}_t)\right\}$ (similar expressions for the other sources).
Then for given positive numbers D_X, $D_1, \ldots, D_n$, a rate *(n+1)*-tuplet R_X, $R_1, \ldots, R_n$,, is admissible if for any $\delta > 0$, $n \ge n_0(\delta)$, there exists a *(n+2)-tuplet* $(\varphi_X, \varphi_1, \cdots, \varphi_n, \psi) \in \Im_{n,\delta}(R_x, R_1, \cdots, R_n)$, such that $\Delta_i \le D_i + \delta$.

In our specific case, we do not care about the reproduction of the Y_i's, so the D_i's can be large. Rather, the Y_i's act as helpers to reproduce X by providing side information at the data fusion node. This is the so called *n*-helper case. Then for an encoding system using the Y_i,'s as *n*-helpers, the rate-distortion region given by:
$\Re(D_x, D_i, \cdots D_n) = \{(R_x, R_i, \cdots R_n) : (R_x, R_i, \cdots R_n) \text{ is admissible}\}$ for a given set of rates and distortion measures, is desired.

Main Result:

For the special case of the correlated Gaussian source, we can now state the main result that we have obtained. We consider the encoding functions:
$\varphi_X : X^m \to \aleph_1 = \{1, \cdots, C_1\}, \cdots, \varphi_i : Y_i^m \to \aleph_i = \{1, \cdots, C_i\}$ to be such that the rate constraints being satisfied are: $\frac{1}{m}\log C_i \le R_i + \delta$, *i=X,1, 2, ...n.* Extending previous results [4]-[10], we show that for an admissible rate $(R_X, R_1, R_2, \cdots R_n)$, and for some D_i's > 0, the *n*-helper system data rates for correlated Gaussian sources can be fused to yield an effective date rate (with respect to source *X*) satisfying the following lower bound:

$$R_X \ge \frac{1}{2}\log\left\{\frac{\sigma_X^2}{D_X}\cdot\left[\prod_{k=1}^{n}\left(1-\rho_{XY_k}^2+\rho_{XY_k}^2\cdot 2^{-2R_k}\right)\right]^{\frac{1}{n}}\right\}$$

This is the desired rate distortion region.

Proof:

In the interests of brevity, an outline of the main steps of the proof is given. The method of employing joint weakly δ-typical tuplet, based on typical sequences, is used in the proofs of the characterization, rather than a measure-theoretic approach [11].

We assume that an admissible set of rates exists, and we let $W_x = \varphi_X\left(X^n\right), W_i = \varphi_i\left(Y^n\right)$.
Then:
$n(R_x+\delta) \ge \log(C_1) \ge H(W_1) \ge I\left(X^n;\hat{X}^n\right) - \frac{1}{n}\sum_{k=1}^{n} I\left(X^n;W_k\right)$ and $\quad n(R_i+\delta) \ge \log(C_i) \ge H(W_i) \ge I\left(Y_i^n;W_i\right)$.
Note that $W_i \to Y_i^n \to X^n$ forms a Markov Chain, thus defining:

adjust their constellation phase with respect to each other so that their signals arrive at the receiver in phase. This is called RF synchronization.

It is also possible that the channel state information is not fed back to the transmitter side. Thus, the constellation phase cannot be controlled, and RF synchronization cannot be achieved. Signals may arrive at the receivers constructively or destructively. Individual transmitted signals must be determined at the receiver with the help of channel state information. Since the constellation and channel state information are known, the possible combined receive signal can be computed, and coherent communication schemes can be used if the phase of the combined receive signal can be tracked as well.

When channel state information cannot be estimated, obviously, there is no way to achieve RF synchronization. Since the possible clean received signal cannot be determined, it would be more feasible to adopt noncoherent communication schemes.

V. CAPACITY RESULTS

The information theoretic capacity for multiple antennas communications has been investigated to by Foschini[16], and Telatar [17] independently. Let us assume there are n_T transmitters and n_R receivers, and there is no feedback link from transmitter to receivers. The n_T transmitted signal components are taken to be statistically independent Gaussians. The virtue of a Gaussian distribution for the transmitted signal is well established, see [18]. Under the assumption of uncorrelated noise on each channel, the capacity takes the form [18-20],

$$C = \log_2 \det(\mathbf{I}_{n_r} + \frac{1}{No}\mathbf{G}\mathbf{A}_s\mathbf{G}^H) \quad (2)$$

where $\mathbf{A}_s$ is the transmitted signal covariance matrix, $\mathbf{G}$ is the channel response matrix that includes path loss and fading, and No is the noise variance. It is observed that capacity is a function of the channel state information $\mathbf{G}$ and the transmitted signal covariance matrix $\mathbf{A}_s$. Furthermore, from Hadamard's inequality [21], in order to maximize the capacity, $\mathbf{G}\mathbf{A}_s\mathbf{G}^H$ must be a diagonal matrix. When channels are independent from each other, $\mathbf{G}$ and $\mathbf{G}^H$ are diagonal, and $\mathbf{A}_s$ has to be diagonal as well.

For Gaussian channels, it is obvious to see the best power allocation policy is the equal power distribution. Several combinations of the numbers of transmitters and receivers are briefly discussed here.

1. Receive Diversity: maximum ratio combining, $n_T=1$, $n_R=n$
For Gaussian channels,

$$C = \log_2[1 + SNR * n] \quad (3)$$

For unequal path loss fading channels,

$$C = \log_2[1 + SNR * \|\mathbf{G}\|^2] \quad (4)$$

where $\mathbf{G}$ is a nx1 vector, and SNR is defined as the ratio of transmit power to the receiver noise power.

2. Transmit Diversity $n_T=n$, $n_R=1$.
The capacity can be derived from (2). For Gaussian channels,

$$C = \log_2[1 + \sum_i SNR_i] \quad (5)$$

where SNR_i is defined as the ratio of the i-th transmit power to the receive noise power, assuming the noise power is equal at all receivers. For unequal path loss channels, increase for the data rate. Thus, in order to deal with a peak power constraint, increased communications range or rate can be achieved if the transmitters first share the message to transmit, and then all send it. Some weighted combining scheme can be used in the receiver array. Using more bandwidth as the number of transmitters is increased is thus a simple way to improve energy efficiency, taking advantage of the absence of competing uses.

For fading channels, optimal combining is simple as well. Because each channel is uncorrelated, the number of received signals are the number of transmitters times the number of receivers. The same optimal combining formula applies to these received signals.

In addition to the bandwidth expansion, it is noted that control signaling overhead is necessary. Its function is to allocate and manage channel resources as well as ensure the messages of the transmitters are coordinated. For FDMA and

CDMA, the receiver complexity grows with the numbber of transmitters linearly. For TDMA, the receiver speed goes up with the number of transmitters. For multiuser OFDM, the number of transmitters doesn't affect the receiver complexity so dramatically. However, time and frequency synchronization [12][13] are critical issues that must be dealt with carefully.

Multiplexing affects the transmitter design as well. For FDMA and CDMA, each transmitter uses a unique frequency band and spreading code, respectively. This uniqueness could be implemented by hardware or by software. TDMA and multiuser

OFDM transmitters are assigned channel resources dynamically, and this results in more complicated hardware design.

No matter whether multiplexing is adopted, it is valuable to investigate communications with multiple transmitters over the same channel. They are orthogonal communication schemes in the sense that multiplexing just deals with how channels are allocated, and doesn't care how each channel is used. So, there could be different coding, modulation, etc. at different multiplex channels. In our study, we address the communications over the same channel. One advantage of this kind of communications is the simplicity of transmitter design. Transmitters using the same channel can be almost identical.

We consider two kinds of channels that may be involved. They are the AWGN channel and the fading channel with unequal path loss. The availability of channel state information affects RF synchronization and the choice of modulation and demodulation schemes. It is easy to appreciate the importance of estimating the channel state information for fading channels. It includes the magnitude and phase of the channel response, both critical quantities in optimal combining. For AWGN channels, there is no path loss. However, due to the different path lengths, transmitted signals arrive at the transmitter with different RF phases. This phase relationship is part of the channel state information. Here, acknowledge of channel state information is characterized into three types, as follows.

We assume channels are not reciprocal. Thus, channel state information must be estimated at the receiver side. One approach is to insert a pilot sequence [14][15] periodically in each transmitted signal sequence. Because more than one transmitter uses the same channel and there is interference between pilot sequences from different transmitters, some properties must be satisfied among these pilot sequences so that the channel state information for each channel can be estimated. Then, the channel state information is feed back to the transmitters explicitly or implicitly. From this channel state information, transmitters can

$$C = \log_2[1 + \sum_i SNR_i * \| \mathbf{G}_i \|^2] \quad (6)$$

where $\mathbf{G}_i$ is the channel response from the i-th transmitter to the single receiver. It has a chi-square distribution with 2 degrees of freedom and a variance that depends on the path loss.

3. Combined Transmit-Receive Diversity.

The capacity can be computed numerically by Monte Carlo simulation. For $n_T \geq n_R$, equal transmit power and equal path loss Rayleigh fading channels, the capacity lower bound was derived in [17]

$$C > \sum_{k=n_T-(n_R-1)}^{n_T} \log_2[1 + \chi_{2k}^2 SNR] \quad (7)$$

where χ_{2k}^2 is a chi-square distribution with 2k degrees of freedom.

The above discussion for transmit diversity is based on the situation when RF phases are not synchronized at the receivers. When their relative phases are adjusted so that they are aligned at the receiver, the amplitude of the combined signal is the sum of the constituent transmit amplitudes. Therefore, for unequal path loss fading channels, the capacity is

$$C = \log_2 \left[1 + \left(\sum_{k=1}^{n_T} \sqrt{SNR_i * \| \mathbf{G}_i \|^2} \right)^2 \right] \quad (8)$$

The capacity for Gaussian channels is obtained by replacing $\mathbf{G}_i$ with 1. For receiver diversity, RF synchronization is not a problem because there is only one transmit signal at any instant of time. For combined transmit-receive diversity, RF synchronization is almost impossible to be achieved simultaneously at all receivers unless transmitters and receivers are deliberately located in the fixed channel environment, and can move to any location in the varying channel environment. The capacity for RF synchronization is much higher than that for RF non-synchronization.

VI. A PRACTICAL APPROACH

In wireless sensor networks, the number of available transmitters may vary with time. Sensor failures or additions are expected to take place. Thus, the communication scheme used in the environment must be adaptive with respect to the number of transmitters. Space-time trellis codes [22] are designed for fixed numbers of transmitters. We propose an approach which is easier to implement and more insensitive to the change of the number of transmitters. In the approach, the relative constellation phases of transmitters change every symbol interval so that signals combined destructively may combine constructively in other symbol intervals. Consequently, the channel behaves like a fast fading channel. We have conducted a computer search for the best relative phase rotation. There are two transmit antennas using BPSK signal constellations and the rate 1/2 repetition code. The channels are AWGN channels. The constellation phase of the first transmit antenna is fixed, while the second transmit antenna increases its constellation phase by the value of relative phase rotation at the second symbol interval. It turns out that the best phase is 180 degrees. In other words, the second transmit antenna flips its constellation phase every symbol interval. It is interesting that the BER at this point is equal to the BER for one transmit antenna with the same total transmit power. If the constraint is on the individual transmit power, which is the usual case in wireless sensor network, we will obtain 3dB gain in performance. For three transmit antennas, the minimum appears at three points, $(0,\pi)$, (π,π) and $(\pi,0)$. These three points can be obtained with the same operation. Two constellations change by the same phase, while the other changes 180 degree more. When 2 and 3 transmit antenna cases are considered together, we can observe the robustness to the number of transmit antennas. When one transmit antenna is added to the 2 transmit antenna scenario, the original two antennas work in the same way, while the third antenna can choose to flip every symbol interval, as the second antenna, or not, as the first antenna. Therefore, the system scales up without changing the operation of the existing transmitters.

VII. CONCLUSION

In this work, we have considered the rate distortion problem for a sensor network employing data fusion at a node. The main assumption has been that the n-helper nodes are all producing correlated Gaussian data, which has then enabled us to obtain an analytic form for the rate distortion bound. The primary utility of this result is to compare practical data fusion schemes with the predicted bounds—in particular, to determine which are the most critical and sensitive parameters affecting rate and performance of a data fusion scheme in a network sensor system.

The main limitation of the work is that currently, extensions to the more 'real-world' scenario of non-Gaussian sources and channels are not obvious. However, efforts are underway to formulate numerically solvable versions for the more realistic scenarios. Our ultimate goal is, in the absence of any tractable analytical expression, to obtain at least an iterative algorithm, or a convex optimization form. Ideally, this would allow us to accurately predict maximum rate/ minimum distortion pairs for a wide variety of channels and sources for random arrangements of sensors. We would also like to be able to investigate various 'what-if' scenarios in simulation set-ups for particular types of configurations and coding implementations. Thus, we hope eventually to be able to use these bounds to definitively compare various data fusion and network communication schemes for wireless sensor networks, with regards to their performance and efficiency.

We have also presented a simple phase rotation scheme that enables a variable number of transmitters without a common phase reference to achieve gains over simple non-coherent combining, assuming the same channel must be used. However, much work remains to be done in devising schemes with superior performance. Simple capacity calculations indicate that substantial gains can be obtained by providing a coherent phase reference, and thus it is likely that approaches which devote a significant fraction of the energy to beacons may be fruitful.

REFERENCES

[1] P.K. Varshney, *Distributed Detection and Data Fusion*, New York: Springer-Verlag, 1997.

[2] G. J. Pottie, et. al., "Wireless Sensor Networks", *Information Theory Workshop Proceedings, 1998*, Killamey, Ireland, June 22-26, 1998.

[3] G. J. Pottie, "Multi-node Processing Problems in Distributed Sensor Networks", *IEEE International Symposium on Information Theory*, Cambridge, MA Aug 16-21, 1998.

[4] El Gamal and T. M. Cover, "Multiple user information theory", *Proceedings of the IEEE*, 68: 1466-1483, 1980.

[5] K. Sohrabi, G. Pottie, "Performance of a self-organizing algorithm for wireless ad-hoc sensor networks", *IEEE Vehicular Technology Conference*, Fall 1999.

[6] D. Slepian and J. K. Wolf, “Noiseless coding of correlated information sources”, *IEEE Transactions on Information Theory*, vol. IT-19, pp. 471-480, July 1973.

[7] Y. Oohama, “Gaussian multi-terminal source coding”, *IEEE Transactions on Information Theory*, vol. 43, no. 6, Nov. 1997.

[8] D. Wyner and J. Ziv, “The rate-distortion function for source coding with side information at the decoder”, *IEEE Transactions on Information Theory*, vol. IT-22, pp. 1-10, Jan. 1976.

[9] T. S. Han and K. Kobayashi, “A unified achievable rate region for a general class of multi-terminal source coding systems”, *IEEE Transactions on Information Theory*, vol. IT-26, pp. 277-288, May 1980.

[10] T. S. Han, “Hypothesis testing with multi-terminal data compression”, *IEEE Transactions on Information Theory*, vol. IT-33, pp. 759-772, Nov. 1987.

[11] Csiszar and J. Korner, “Towards a general theory of source networks”, *IEEE Transactions on Information Theory*, vol. IT-26, pp. 155-165, Mar. 1980.

[12] L. Wei and C. Schlegel, "Synchronization requirements for multi-user OFDM on satellite mobile and two-path Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 43, pp. 887-895, 1995.

[13] J. J. van de Beek, P. O. Borjesson, M. L. Boucheret, D. Landstrom, J. M. Arenas, P. Odling, C. Ostberg, M. Wahlqvist, and S. K. Wilson, "A time and frequency synchronization scheme for multiuser OFDM," *IEEE J. Select. Areas Commun.*, vol. 17, pp. 1900-1914, Nov. 1999.

[14] A. F. Naguib, V. Tarokh, N. Seshadri, and A. R. Calderbank, "A space-time coding modem for high-data-rate wireless communications," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 1459-1478, Oct. 1998.

[15] J.-C. Guey, M. P. Fitz, M. R. Bell, and W.-Y. Kuo, "Signal design for transmitter diversity wireless communication systems over Rayleigh fading channels," *IEEE Trans. Commun.*, vol. 47, pp. 527-537, Apr. 1999.

[16] G. J. Foschini, Jr. and M. J. Gans, "On limits of wireless communication in a fading environment when using multiple antennas," *Wireless Personal Commun.*, vol. 6, no. 3, pp. 311-335, Mar. 1998.

[17] E. Telatar, "Capacity of multiantenna Gaussian channels," AT&T-Bell Lab, Internal Tech. Memo., June 1995

[18] M. S. Pinsker, *Information and Information Stability of Random Processes*. San Francisco: Holden Bay, Chapter 10, 1964.

[19] S. Kullback, Information Theory and Statistics. New York: John Wiley and Sons, 1959.

[20] D. B. Osteyee, and I. J. Good, Information Weight of Evidence, the Singularity between Probability Measures and Signal Detection. New York: Springer-Verlag, 1970.

[21] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley and Sons, 1991.

[22] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: Performance analysis and code construction," *IEEE Trans. Inform. Theory*, vol. 44, pp. 744-765, Mar. 1998.

An Adaptive Multi-Queue Service Room Protocol for Wireless Networks with Multipacket Reception

Qing Zhao and Lang Tong
School of Electrical and Computer Engineering
Cornell University
Ithaca, NY 14853
{qzhao,ltong}@ee.cornell.edu

Abstract

An adaptive medium access control (MAC) protocol for heterogeneous networks with finite population is proposed. Referred to as the Multi-Queue Service Room (MQSR) protocol, this scheme is capable of handling users with different Quality-of-Service (QoS) constraints. By exploiting the multipacket reception (MPR) capability, the MQSR protocol adaptively grants access to the MPR channel to a number of users such that the expected number of successfully received packets is maximized in each slot. The optimal access protocol avoids unnecessary empty slots for light traffic and excessive collisions for heavy traffic. It has superior throughput and delay performance as compared to, for example, the slotted Aloha with the optimal retransmission probability. This protocol can be applied to random access networks with multimedia traffic.

1 Introduction

In multiaccess wireless networks where a common channel is shared by a population of users, a key issue, referred to as medium access control (MAC), is to coordinate the transmissions of all users so that the common channel is efficiently utilized and the Quality-of-Service (QoS) requirement of each user is satisfied. The schemes for coordinating transmissions are called MAC protocols.

Based on the assumption of homogeneous QoS constraints and a collision channel model where concurrent transmission of two or more packets results in the destruction of all the transmitted information, numerous MAC protocols, such as Aloha [1, 13], tree algorithm [4], first come first serve algorithm [6], and a class of adaptive schemes [12, 3, 10, 9], have been proposed. However, with the development of spread spectrum, space-time coding, and new signal processing techniques, this collision channel model does not hold in many important practical communication systems where one or more packets can be successfully received in the presence of other simultaneous transmissions. For instance, the capture phenomenon is

---

*This work was supported in part by the National Science Foundation under Contract CCR-9804019 and the Multidisciplinary University Research Initiative (MURI) under the Office of Naval Research Contract N00014-00-1-0564.

common in local area radio networks. Other examples include networks using CDMA and/or antenna array along with multiuser detection techniques.

This new channel model which offers the capability of multipacket reception (MPR) presents a new challenge for the design of MAC protocols. Some researchers have considered the issue of extending existing MAC protocols to channels with MPR capability. In [5, 11], the contention free scheme TDMA is extended to a fully connected half-duplex ad hoc network with total P conventional collision channels. In [7, 8], the performance of slotted Aloha for MPR channels with infinite population is analyzed. In [14], the authors propose a dynamic queue protocol designed explicitly for channels with MPR capability. However, these protocols still assume homogeneous QoS constraints among all users, which may not hold in multimedia networks.

In this paper, we propose a MAC protocol for networks with MPR capability and heterogeneous QoS requirements. In particular, we consider a finite population of users whose QoS requirement is characterized by the packet delay at the heaviest traffic load. Since, in general, packet delay increases with the traffic load, this delay constraint specifies the worst case performance of the network. Our goal is to design a MAC protocol that maximizes the expected number of successfully received packets in each slot while ensuring each user's QoS requirement. To satisfy the heterogeneous delay constraints, users are assigned with different priorities for accessing the channel. According to their priorities, an appropriate subset of users who gain access to the channel are chosen at the beginning of each slot so that the expected number of successfully received packets is maximized in each slot. As a consequence, significant improvement in throughput and delay performance over existing MAC schemes such as slotted Aloha with optimal retransmission probability is achieved by the proposed protocol.

2 The Problem Statements

2.1 The Model

We consider a communication network with M users who transmit data to a central controller through a common channel. Each user generates data in the form of equal-sized packets. Transmission time is slotted and each packet requires one time slot to transmit. Each user has a single buffer. At the beginning of each slot, a user independently generates a packet with probability p, but only accepts this packet if its buffer is currently empty. A packet generated at the beginning of a slot may be transmitted in this slot, and a successfully transmitted packet leaves its buffer. Packets generated by a user with a full buffer are assumed lost.

Users are partitioned into L groups according to their QoS constraints. The M_l $(l = 1, \cdots, L, \sum_{l=1}^{L} M_l = M)$ users in the lth group require their packet delay at $p = 1$ no greater than d_l, where we define packet delay as the expected number of slots from the time a packet enters a buffer until the end of its successful transmission.

The slotted channel is such that the probability of having k successes in a slot where there are n transmissions depends only on the number of transmitted packets

$$C_{n,k} = \mathrm{P}[k \text{ packets are correctly received} \mid n \text{ are transmitted}] \quad (1 \leq n \leq M, 0 \leq k \leq n).$$

The multipacket reception matrix of the channel is then defined as

$$\mathbf{C} = \begin{pmatrix} C_{1,0} & C_{1,1} & & & \\ C_{2,0} & C_{2,1} & C_{2,2} & & \\ \vdots & \vdots & \vdots & & \\ C_{M,0} & C_{M,1} & C_{M,2} & \cdots & C_{M,M} \end{pmatrix}. \tag{1}$$

For such an MPR channel, we define the channel capacity as

$$\eta \triangleq \max_{n=1,\cdots,M} \mathcal{C}_n, \tag{2}$$

where

$$\mathcal{C}_n \triangleq \sum_{k=1}^{n} kC_{n,k} \tag{3}$$

is the expected number of packets correctly received when n packets are transmitted. Let

$$n_0 \triangleq \min\{\arg \max_{n=1,\cdots,M} \mathcal{C}_n\}. \tag{4}$$

We can see that at heavy traffic load, n_0 packets should be transmitted simultaneously to achieve the channel capacity η. Noticing that the number of simultaneously transmitted packets for achieving η may not be unique, we define n_0 as the minimum to save transmission power. For MPR channels with n_0 greater than 1, contention should be preferred at any traffic load in order to fully exploit the channel MPR capability.

This general model for MPR channels, also considered in [7, 8, 2], applies to the conventional collision channel and channels with capture as special examples. The reception matrix of the conventional collision channel and channels with capture are given by

$$\begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 1 & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & 0 & 0 & \cdots & 0 \end{pmatrix}, \quad \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 1-p_2 & p_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 1-p_M & p_M & 0 & \cdots & 0 \end{pmatrix}, \tag{5}$$

where p_i is the probability of capture given i simultaneous transmissions.

We assume that the central controller can distinguish without error between empty and nonempty slots. Furthermore, if some packets are successfully demodulated at the end of a slot, the central controller can identify the source of these packets and inform their sources so that their buffers can be released. However, if at least one packet is successfully demodulated at the end of a slot, the central controller does not assume the knowledge whether there are other packets transmitted in this slot but not successfully received. We illustrate this point in Figure 1 where we consider possible outcomes of a slot: empty, nonempty with success, and nonempty without success (successfully received packets are illustrated by shaded rectangles). To the central controller, the two events happened in the third and forth slots are indistinguishable.

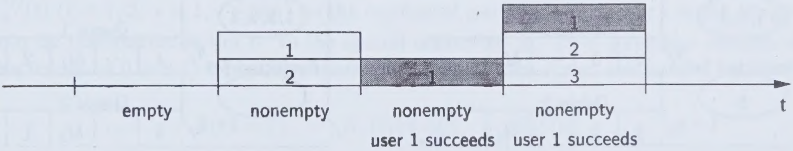


Figure 1: Possible Outcomes of A Slot

2.2 The Problem

Our goal here is to design a MAC protocol for an MPR channel so that the expected number of successfully received packets in each slot is maximized under the constraint that the packet delay D_l of the lth group at $p = 1$ is no larger than d_l for $l = 1, \cdots, L$. Before pursuing the protocol design, we give the necessary and sufficient condition for the existence of a MAC protocol with such delay constraints.

Proposition 1 *Let M_l $(l = 1, \cdots, L)$ be the number of users who require their packet delay at $p = 1$ no larger than d_l. Then there exists a MAC protocol that guarantees each user's delay requirement if and only if*

$$\sum_{l=1}^{L} \frac{M_l}{d_l} \leq \eta, \tag{6}$$

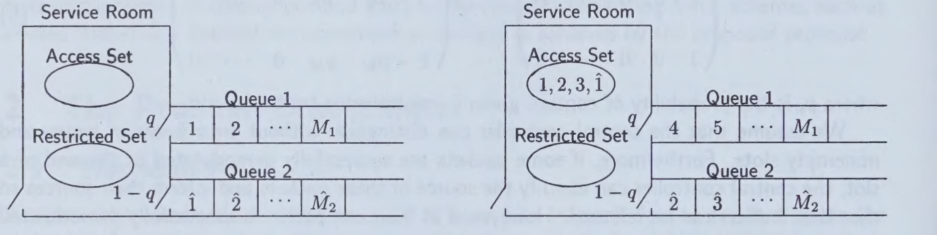
where η is the channel capacity.

3 The Multi-Queue Service Room Protocol

3.1 The Basic Structure

We develop the multi-queue service room (MQSR) protocol for the case of $L = 2$, where users in the first group require $D_1 \leq d_1$. Its extension to cases with $L > 2$ is straightforward.

We illustrate the basic procedure of the MQSR protocol in Figure 2, where users from the second group are indicated by $\hat{i}$ $(i = 1, \cdots, M_2)$. As shown in Figure 2-A, users of the two groups are waiting, respectively, in two queues to enter the service room. Users in the service

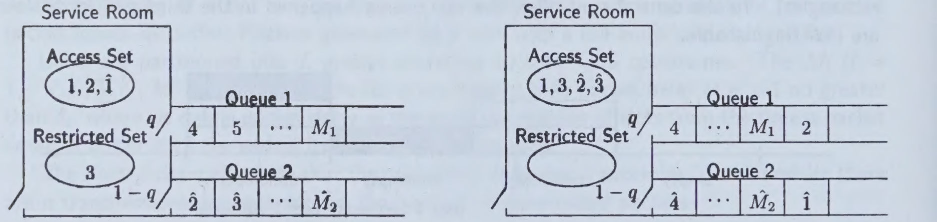

Service Room
Access Set
Restricted Set
Queue 1
Queue 2
q
$1-q$
1
2
$\cdots$
M_1
$\hat{1}$
$\hat{2}$
$\hat{M}_2$
Service Room
Access Set
$1, 2, 3, \hat{1}$
Restricted Set
Queue 1
Queue 2
q
$1-q$
4
5
$\cdots$
M_1
$\hat{2}$
$\hat{3}$
$\hat{M}_2$

A.

B. $K(1) = 4$

$K_1(1) \sim B(q, K) = 3,\ K_2(1) = K(1) - K_1(1) =$

Outcome: nonempty, no one succeeds; $\mathcal{P}(1) = \phi$


Service Room
Access Set
$1, 2, \hat{1}$
Restricted Set
3
Queue 1
Queue 2
q
$1-q$
4
5
$\cdots$
M_1
$\hat{2}$
$\hat{3}$
$\hat{M}_2$
Service Room
Access Set
$1, 3, \hat{2}, \hat{3}$
Restricted Set
Queue 1
Queue 2
q
$1-q$
4
$\cdots$
M_1
2
$\hat{4}$
$\hat{M}_2$
$\hat{1}$

C. $K(2) = 3$

$K_1(2) \sim B(q, K) = 2,\ K_2(2) = K(2) - K_1(2) = 1$

Outcome: user 2, $\hat{1}$ succeed; $\mathcal{P}(2) = \{2, \hat{1}\}$

D. $K(3) = 4$

$K_1(3) \sim B(q, K) = 2, K_2(3) = K(3) - K_1(3) =$

Outcome: empty; $\mathcal{P}(3) = \{1, 3, \hat{2}, \hat{3}\}$

Figure 2: The Basic Procedure of the Multi-Queue Service Room Protocol.

room may transmit packets generated before they enter the service room. Packets generated by a user inside the service room are held in the user's buffer (if the buffer is empty) and can not be transmitted until the next time this user enters the service room. After entering the service room, a user stays there until the central controller detects that this user has been processed, *i.e.*, it has no packet to transmit. At this time, this user leaves the service room and goes to the end of its queue.

At the end of slot t, the central controller first determines the set, denoted by $\mathcal{P}(t)$, of processed users of slot t. After removing the processed users from the service room, the central controller decides the size $K(t+1)$ of the access set which consists of users who gain access to the channel in slot $t+1$. With probability q, a user who joins the access set is from the first group, and with probability $1-q$, it is from the second group. Let $K_l(t+1)$ $(l=1,2)$ be the number of users from the lth group who will access the channel in slot $t+1$. Then $K_1(t+1)$ obeys a binomial distribution with $K(t+1)$ trials and a success probability q (denoted by $K_1(t+1) \sim B(q, K(t+1))$), and $K_2(t+1) = K(t+1) - K_1(t+1)$. Let $\alpha_l(t)$ $(l=1,2)$ be the number of users from the lth group who remain in the service room after processed users have been removed from the service room at the end of slot t. Then, if $K_l(t+1) > \alpha_l(t)$, the first $K_l(t+1) - \alpha_l(t)$ users in Queue l enter the service room and join the access set at the beginning of slot $t+1$ (see Figure 2-B and D). On the other hand, if $K_l(t+1) < \alpha_l(t)$, the last (according to their time of entering the service room) $\alpha_l(t) - K_l(t+1)$ users from the lth group join the restricted set at the beginning of slot $t+1$ (see Figure 2-C). Users in the restricted set remain in the service room, but they can not access the channel in current slot.

After specifying the basic structure of the MQSR protocol, we now consider parameters that remain to be designed. The first parameter to be determined is q, an indicator of the priority of users in the first group over users in the second group. Since q is constant for each slot, it can be designed off line. The two parameters to be determined on line are $K(t)$, the size of the access set for slot t, and $\mathcal{P}(t)$, the processed set of slot t. The problem of choosing q and determining $K(t)$ for each t is formulated in Section 3.2 and the determination of $\mathcal{P}(t)$ is detailed in Section 3.4.

We point out that the optimal window protocol proposed in [10] has a similar structure as the MQSR protocol proposed here. Relying on exhaustive search, however, the window protocol is only computationally feasible for networks with 2 or 3 users and no MPR. Furthermore, homogeneous QoS constraints are assumed by this protocol.

3.2 Problem Formulation

Let $X_i^{(l)}(t)$ $(l=1,2,\ i=1,\cdots,M_l)$ be the number of packets held by user i in the lth group that may be transmitted in slot t. To the central controller, $X_i^{(l)}(t)$ is a random variable with possible values of 0 and 1. Let $\mathcal{A}(t)$ denote the access set for slot t. We have

$$\mathcal{A}(t) = \{1, \cdots, K_1(t)\} \bigcup \{\hat{1}, \cdots, \hat{K}_2(t)\}, \tag{7}$$

where $\hat{i}$ indicates the ith user of the second group and we assume users of the lth group are relabeled at the beginning of each slot, starting from the service room to the end of the lth queue. Let $F(t)$ be the outcome of slot t, which contains information on whether slot t is empty and whose packets are successfully received in slot t (see Figure 1). Denoted by $I_{[1,t-1]}$, the information available for determining $K(t)$ at the beginning of slot t is the initial condition of the network in the form of the distribution of $X_i^{(l)}(1)$ $(l=1,2,\ i=1,\cdots,M_l)$, the access sets $\mathcal{A}(1), \cdots, \mathcal{A}(t-1)$ for previous slots, and the outcomes $F(1), \cdots, F(t-1)$ of previous slots.

Finally, let $S(t)$ denote the number of successfully transmitted packets in slot t. It is a random variable whose distribution conditioned on $I_{[1,t-1]}$ depends on $\mathcal{A}(t)$ and the channel MPR matrix $\mathbf{C}$.

We are now ready to formulate the problem of determining q and $K(t)$ for each t.

Given $\mathbf{C}$ *and* p*, determine* q *and* $K(t)$ *for each* t *based on* $I_{[1,t-1]}$ *by solving the following constrained optimization problem:*

$$\{q, K(t)\} = \arg \max_{\substack{K(t)=1,\cdots,M,\\ K_1(t)\sim B(q,K(t)),\\ K_2(t)=K(t)-K_1(t)}} E[S(t) \mid (K_1(t), K_2(t)),\ I_{[1,t-1]}] \ \textit{subject to}\ D_1 \leq d_1, \tag{8}$$

where $K_1(t)$ *is a realization of the binomial distribution with parameters* q *and* $K(t)$*, and* $E[S(t) \mid (K_1(t), K_2(t)),\ I_{[1,t-1]}]$ *is the expected number of successfully transmitted packets given* $I_{[1,t-1]}$ *and* $\mathcal{A}(t)$ *which consists of the first* $K_l(t)$ $(l=1,2)$ *users of the* l*th group.*

This constrained optimization problem can be decoupled into two steps. We first choose q so that the delay constraint $D_1 \leq d_1$ is satisfied. Then with q determined, choose $K(t)$ for each t so that $E[S(t) \mid I_{[1,t-1]}]$ is maximized. These two steps are detailed in, respectively, Section 3.3 and Section 3.5.

3.3 The Determination of q

At $p=1$, every user has a packet to transmit at the beginning of each slot. To maximize $E[S(t) \mid I_{[1,t-1]}]$, we should choose $K(t)=n_0$ for each t. In this case, the throughput T, defined as the expected number of successfully transmitted packets in one slot, achieves the channel capacity η. Let T_1 denote the throughput for users in the first group, *i.e.*, the expected number of packets from the first group that are successfully received in one slot. We can show that q is the ratio of T_1 to the overall throughput η, *i.e.*,

$$q = \frac{T_1}{\eta}. \tag{9}$$

Furthermore, for a network where users have homogeneous and independent traffic load, we have the following relation between throughput and delay under equilibrium condition [10]:

$$D_1 = 1 + \frac{M_1}{T_1} - \frac{1}{p}. \tag{10}$$

Since $p=1$ in our definition of D_1 and T_1, we have

$$D_1 = \frac{M_1}{T_1}. \tag{11}$$

Thus, $D_1 \leq d_1$ implies $T_1 \geq M_1/d_1$. From (9) we then have

$$q \geq \frac{M_1}{d_1\eta}. \tag{12}$$

To avoid the second group making unnecessary sacrifice, we choose

$$q = \frac{M_1}{d_1\eta}. \tag{13}$$

3.4 The Determination of the Processed Set

At the end of slot t, the central controller needs to determine the processed set $\mathcal{P}(t)$. By definition, $\mathcal{P}(t)$ is a subset of $\mathcal{A}(t)$. It consists of users whose packets are successfully transmitted in slot t and users who the central controller can assert have an empty buffer upon entering the service room. Let $X_i^{(l)}(t^+)$ $(l = 1, 2, i = 1, \cdots, K_l(t))$ denote the number of packets held by the ith user of the lth group at the end of slot t after successfully transmitted packets have been removed from their buffer. We then have

$$\begin{aligned} \mathcal{P}(t) = \quad & \{i:\ 1 \le i \le K_1(t),\ E[X_i^{(1)}(t^+) \mid I_{[1,t]}] = 0\} \\ & \bigcup\{\hat{i}:\ 1 \le i \le K_2(t),\ E[X_i^{(2)}(t^+) \mid I_{[1,t]}] = 0\}. \end{aligned} \tag{14}$$

We now consider evaluating $E[X_i^{(l)}(t^+) \mid I_{[1,t]}]$. If slot t is empty, we readily have $E[X_i^{(l)}(t^+) \mid I_{[1,t]}] = 0$ for $l = 1, 2$ and $i = 1, \cdots, K_l(t)$, and

$$\mathcal{P}(t) = \mathcal{A}(t). \tag{15}$$

If, on the other hand, slot t is nonempty and $s_i^{(l)}$ packets from the ith user of the lth group are successfully received at the end of slot t, we have $E[X_i^{(l)}(t^+) \mid I_{[1,t]}] = 0$ for $(i, l) \in \{(i, l) : s_i^{(l)} = 1\}$. However, it is also possible that $E[X_i^{(l)}(t^+) \mid I_{[1,t]}] = 0$ for some (i, l) with $s_i^{(l)} = 0$ (consider, as an extreme example, the conventional channel; one successful transmission implies that other users in the access set do not have packets). In order to evaluate $E[X_i^{(l)}(t^+) \mid I_{[1,t]}]$ for $\{(i, l) :\ s_i^{(l)} = 0\}$, we compute the conditional joint distribution of $\{X_i^{(l)}(t^+),\ l = 1, 2,\ i = 1, \cdots, N_l(t)\}$, where $N_l(t)$ is the number of users from the lth group that are inside the service room in slot t. For $0 \le x_i^{(l)} \le 1 - s_i^{(l)}$ $(l = 1, 2,\ i = 1, \cdots, N_l(t))$, we have,

$$\begin{aligned} & P[\{X_i^{(l)}(t^+) = x_i^{(l)},\ l = 1, 2,\ i = 1, \cdots, N_l(t)\} \mid I_{[1,t]}] \\ & = \frac{P[\{X_i^{(l)}(t^+)=x_i^{(l)},\ l=1,2,\ i=1,\cdots,N_l(t)\},\ F(t) \mid \mathcal{A}(t),\ I_{[1,t-1]}]}{P[F(t) \mid \mathcal{A}(t),\ I_{[1,t-1]}]} \\ & = \frac{C_{z,S} P[\{X_i^{(l)}(t)=x_i^{(l)}+s_i^{(l)},\ l=1,2,\ i=1,\cdots,N_l(t)\} \mid I_{[1,t-1]}]}{\sum_{\{0\le x_i^{(l)} \le 1-s_i^{(l)}, l=1,2,\ i=1,\cdots,N_l(t)\}} C_{z,S} P[\{X_i^{(l)}(t)=x_i^{(l)}+s_i^{(l)},\ l=1,2,\ i=1,\cdots,N_l(t)\} \mid I_{[1,t-1]}]} \end{aligned} \tag{16}$$

where $z = \sum_{l=1}^{2} \sum_{i=1}^{K_l(t)} (x_i^{(l)} + s_i^{(l)})$, and $S = \sum_{l=1}^{2} \sum_{i=1}^{K_l(t)} s_i^{(l)}$ is the total number of successfully transmitted packets in slot t.

With the conditional joint distribution of $\{X_i^{(l)}(t^+),\ l = 1, 2,\ i = 1, \cdots, N_l(t)\}$, we can evaluate $E[X_i^{(l)}(t^+) \mid I_{[1,t]}]$ and obtain $\mathcal{P}(t)$. Recall that $\alpha_l(t)$ denotes the number of unprocessed users from the lth group in slot t, and without loss of generality, we assume they are the first $\alpha_l(t)$ users of the lth group. Since unprocessed users remain in the service room in slot $t + 1$, packets generated by them at the beginning of slot $t + 1$ can not be transmitted until the next time they enter the service room. We thus have $X_i^{(l)}(t + 1) = X_i^{(l)}(t^+)$ for $l = 1, 2,\ i = 1, \cdots, \alpha_l(t)$. Hence, the conditional joint distribution of $\{X_i^{(l)}(t + 1),\ l = 1, 2,\ i = 1, \cdots, \alpha_l(t)\}$ can be easily obtained from the conditional joint distribution of $\{X_i^{(l)}(t^+),\ l = 1, 2,\ i = 1, \cdots, N_l(t)\}$ (given by (16)) by summing out the processed users, *i.e.*,

$$\begin{aligned} & P[\{X_i^{(l)}(t + 1) = x_i^{(l)},\ l = 1, 2, i = 1, \cdots, \alpha_l(t)\} \mid I_{[1,t]}] \\ & = \sum_{\{x_i^{(1)},\ x_j^{(2)},\ i,\hat{j} \in \mathcal{P}(t)\}} P[\{X_i^{(l)}(t^+) = x_i^{(l)},\ l = 1, 2, i = 1, \cdots, N_l(t)\} \mid I_{[1,t]}]. \end{aligned} \tag{17}$$

The joint distribution of $\{X_i^{(l)}(t+1),\ l=1,2,\ i=1,\cdots,\alpha_l(t)\}$ conditioned on $I_{[1,t]}$ serves as the basis for determining $\mathcal{A}(t+1)$ and $\mathcal{P}(t+1)$.

3.5 The Determination of the Access Set

At the end of slot t, after determining the processed users and removing them from the service room, we choose $\mathcal{A}(t+1)$ by specifying $K_1(t+1)$ and $K_2(t+1)$.

For a q given by (13) and all possible sizes of the access set $K(t+1)$ ($K(t+1)=1,\cdots,M$), we choose $(k_1(t+1),k_2(t+1))$, where $k_1(t+1) \sim B(q, K(t+1))$, is a realization of the binomial distribution with parameters q and $K(t+1)$, and $k_2(t+1) = K(t+1) - k_1(t+1)$. Among these M pairs of $(k_1(t+1),k_2(t+1))$, we choose $(K_1(t+1),K_2(t+1))$ so that $E[S(t+1) \mid I_{[1,t]}]$ is maximized, *i.e.*,

$$(K_1(t+1),K_2(t+1)) = \arg \max_{(k_1(t+1),k_2(t+1))} E[S(t+1) \mid (k_1(t+1),k_2(t+1)),\ I_{[1,t]}], \tag{18}$$

where $E[S(t+1) \mid (k_1(t+1),k_2(t+1)),\ I_{[1,t]}]$ denotes the number of successfully transmitted packets in slot $t+1$ if the first $k_l(t+1)$ ($l=1,2$) users from the lth group access the channel. It can be computed as

$$\begin{aligned} &E[S(t+1) \mid (k_1(t+1),k_2(t+1)),\ I_{[1,t]}] \\ &= \textstyle\sum_{n=1}^{k_1(t+1)+k_2(t+1)} C_n P[\sum_{i=1}^{k_1(t+1)} X_i^{(1)}(t+1) + \sum_{i=1}^{k_2(t+1)} X_i^{(2)}(t+1) = n \mid I_{[1,t]}]. \end{aligned} \tag{19}$$

To obtain $P[\sum_{i=1}^{k_1(t+1)} X_i^{(1)}(t+1) + \sum_{i=1}^{k_2(t+1)} X_i^{(2)}(t+1) = n \mid I_{[1,t]}]$ for all possible $(k_1(t+1),k_2(t+1))$, we need the joint distribution of $\{X_i^{(l)}(t+1),\ l=1,2,\ i=1,\cdots,M_l\}$ conditioned on $I_{[1,t]}$. Since users generate packets independently, this conditional joint distribution can be obtained from the conditional joint distribution of $\{X_i^{(l)}(t+1)\ (l=1,2,\ i=1,\cdots,\alpha_l(t))\}$ given by (17) and the marginal distribution of $X_i^{(l)}(t+1)$ ($l=1,2,\ i=\alpha_l(t)+1,\cdots,M_l$). The latter can be computed as

$$P[X_i^{(l)}(t+1)=x] = \begin{cases} (1-p)^{W_i^{(l)}} & \text{if } x=0 \\ 1-(1-p)^{W_i^{(l)}} & \text{if } x=1 \\ 0 & \text{otherwise} \end{cases}, \tag{20}$$

where $W_i^{(l)} = t+1-\tau_i^{(l)}$ with $\tau_i^{(l)}$ defined as the index of the slot in which the ith user in the lth group last time entered the service room or the index of the slot in which this user last time successfully transmitted a packet, whichever is larger.

3.6 Stability

The following theorem gives a sufficient condition under which the proposed protocol is stable.

Theorem 1 *For an MPR channel with* $C_{n,k} > 0$ $(n=1,\cdots,M,\ k=0,\cdots,n)$, *the proposed protocol is stable in the sense that for any* $p \in [0,1]$, *the packet delay of each user is finite.*

It can be shown that the stability of the proposed protocol is equivalent to the finiteness of the expected number of slots a user stays in the service room during one visit, which can be proved with the help of Borel-Cantelli Lemma.

4 Simulation Examples

Presented in this section are simulation studies on the throughput and delay performance of the proposed MQSR protocol in a CDMA network with $M = 10$ users. The MPR capability of this network is provided by 3 orthogonal codes. A packet is transmitted with a code randomly picked from these 3 codes. A packet is successfully received if and only if no other packet transmitted in this slot uses the same code. It can be shown the capacity of this MPR channel is $4/3$.

We first consider the scenario of homogeneous QoS requirement ($L = 1$) and compare the throughput of the proposed MQSR protocol with that of slotted Aloha with optimal retransmission probability. As shown in Figure 3, significant improvement in throughput was achieved by the MQSR protocol. Figure 3 also shows that the MQSR protocol achieved the channel capacity at heavy traffic load.

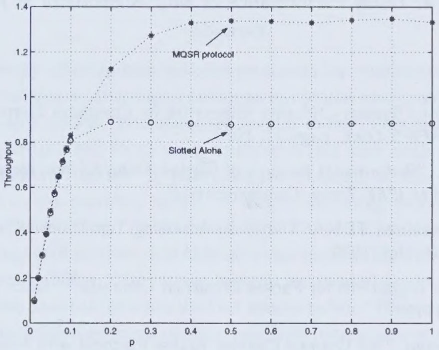


Figure 3: Throughput Comparison between MQSR Protocol and Slotted Aloha with Optimal Retransmission Probability

We now consider the case of $L = 2$, $M_1 = M_2 = 5$, and users of the first group require their packet delay D_1 at $p = 1$ no larger than d_1. We considered different delay requirement of the first group, as illustrated by asterisks in Figure 4. The corresponding q was obtained by (13). The simulated delay of the first group was indicated by the solid line in Figure 4. The circles and dashed line indicate, respectively, the calculated delay and simulated delay of the second group for a given q. Figure 4 shows that the delay requirement of the first group was satisfied for the choice of q given in (13).

5 Conclusion

In this paper, we propose a multi-queue service room MAC protocol designed explicitly for multiaccess networks with MPR capability. The proposed protocol dynamically controls the size of the access set according to the traffic load and the channel MPR capability so that the expected number of successfully transmitted packets is maximized under a set of heterogeneous delay constraints. This protocol achieves the channel capacity at heavy traffic load and is stable in the sense that the packet delay of each user is finite for any traffic load.

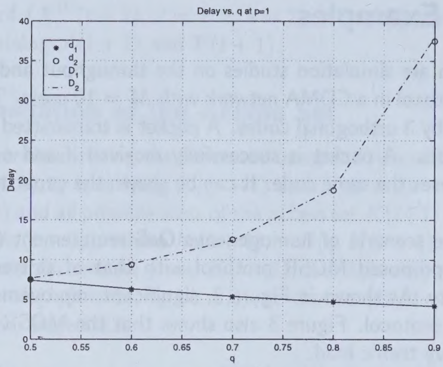


Figure 4: Delay Performance of MQSR Protocol at $p = 1$

References

[1] N. Abramson. "The Aloha System - Another Alternative for Computer Communications". In *Proc. Fall Joint Comput. Conf., AFIPS Conf.*, page 37, 1970.

[2] J. Q. Bao and L. Tong. "Performance Analysis of Slotted Aloha Ad Hoc Networks with Multiple Packet Reception". *Submitted to IEEE Trans. Communications.*

[3] J.I. Capetanakis. "Generalized TDMA: The Multi-Accessing Tree Protocol". *IEEE Trans. Communications*, 27(10):1476–1484, Oct. 1979.

[4] J.I. Capetanakis. "Tree Algorithms for Packet Broadcast Channels.". *IEEE Trans. Information Theory*, 25(5):505–515, Sept. 1979.

[5] I. Chlamtac and A. Farago. "An Optimal Channel Access Protocol with Multiple Reception Capacity". *IEEE Trans. Computers*, 43(4):480–484, April 1994.

[6] R. Gallager. "Conflict Resolution in Random Access Broadcast Networks". In *Proc. of AFOSR Workshop on Comm. Theory and Appl.*, pages 74–76, Sept. 1978.

[7] S. Ghez, S. Verdú, and S.C. Schwartz. "Stability Properties of Slotted Aloha with Multipacket Reception Capability" . *IEEE Trans. Automat. Contr.*, 33(7):640–649, July 1988.

[8] S. Ghez, S. Verdú, and S.C. Schwartz. "Optimal Decentralized Control in the Random Access Multipacket Channel". *IEEE Trans. Automat. Contr.*, 34(11):1153–1163, Nov. 1989.

[9] M.G. Hluchyj. "Multiple Access Window Protocol: Analysis for Large Finite Populations". In *Proc. IEEE Conf. on Decision and Control*, pages 589–595, New York, NY, 1982.

[10] M.G. Hluchyj and R.G. Gallager. "Multiaccess of A Slotted Channel by Finitely Many Users". In *Proc. Nat. Telecomm. Conf.*, pages D4.2.1–D4.2.7, New Orleans, LA., Aug. 1981.

[11] S. Kim and J. Yeo. "Optimal Scheduling in CDMA Pakcet Radio Networks". *Computers and Operations Research*, 25:219–227, March 1998.

[12] L. Kleinrock and Y. Yemini. "An Optimal Adaptive Scheme for Multiple Access Broadcast Communication". In *Proc. International Conference on Communications*, pages 7.2.1–7.2.5, June 1978.

[13] L.G. Roberts. "Aloha Packet System with and without Slots and Capture". In *(ASS Note 8). Stanford, CA: Stanford Research Institute, Advanced Research Projects Agency, Network Information Center.*, 1972.

[14] Q. Zhao and L. Tong. "The Dynamic Queue Protocol for Multiaccess Networks with Multipacket Reception". In *Proc. 34th Asilomar Conf. on Signals, Systems, and Comp.*, Nov. 2000.

Exploiting Distributed Spatial Diversity in Wireless Networks*

J. Nicholas Laneman Gregory W. Wornell
Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139
jnl@allegro.mit.edu

Abstract

We develop energy-efficient transmission protocols for wireless networks that exploit spatial diversity available at distributed antennas to combat multipath fading through coordinated transmission and/or processing by several radios. We discuss the problem in a general network setting and focus on a multiple-access case with sufficient symmetry to make the presentation concise. In particular, we examine several possibilities for the strategy employed by the assisting radios, or relays, including decoding and forwarding and amplifying and forwarding. To characterize performance, we develop outage regions and associated outage probabilities that indicate robustness of the transmissions to varying signal-to-noise ratios (SNRs). The outage regions are treated as set-functions of the observed SNR γ on the channel between the two source radios. The outage regions are thus (conditional) events defined in terms of the other SNR parameters of the channel, and can be readily interpreted in coded and uncoded settings. Our results suggest that the relays employ a threshold rule, namely, for SNRs satisfying $\gamma > \gamma^*$, for some value γ^*, the relays decide to cooperate and pass along each other's transmissions, preferably by decoding and re-encoding with a separate codebook for sufficiently high γ; for SNRs satisfying $\gamma < \gamma^*$, the relays can do better by simply retransmitting their own signals.

1 Introduction

Relaying information over several point-to-point communication links is a basic building block of communication networks. Such relaying is utilized in wired and wireless networks to achieve higher network connectivity (broader coverage), efficient utilization of resources such as power and bandwidth, better economies of scale in the cost of long-haul transmissions (through traffic aggregation), interoperability among networks, and more easily manageable, hierarchical network architectures.

In wireless networks, direct transmission between widely separated radios can be very expensive in terms of transmitted power required for reliable communication. High-power transmissions lead to faster battery drain (shorter network life) as well as increased interference at nearby radios. As alternatives to direct transmission, there are two basic and frequently-employed examples of relayed transmission for wireless networks. In cellular settings, for

---

[*] This work has been supported in part by ARL Federated Labs under Cooperative Agreement No. DAAL01-96-2-0002, and by NSF under Grant No. CCR-9979363.

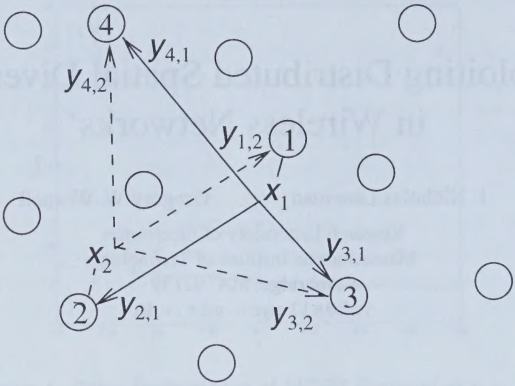


Figure 1: Example wireless network in which transmission protocols for exploiting distributed spatial diversity can be motivated. Indicated are transmitted signals x_i and received signals $y_{j,i}$. Throughout the paper, we focus on transmissions from radios 1 and 2 to radios 3 and 4, respectively.

example, networks provide connectivity between low-power mobiles by providing local connections to high-power basestations that are relayed via a wireline basestation network. In sensor networks, and military battlefield communication networks in general, the use of wireline infrastructure is often precluded and the radios may be substantially power constrained; for these ad-hoc or peer-to-peer networks, transmissions can be relayed wirelessly. As these examples suggest, relayed transmission enlists two or more radios to perform multiple transmissions. The end-to-end transmissions potentially incur higher delay, but because the individual transmissions are over shorter distances (in the wireless case), or over high-quality cabling (in the wireline case), the power requirements for reliable communication can be much lower.

The basic relaying protocols described above are constructed from the sequential use of point-to-point links, where the links are essentially viewed at the network protocol layer; however, more general approaches are possible that involve the coordination of *both* the direct and relayed transmissions, at the network and lower protocol layers, and correspond to scenarios to which the classical relay channel model applies. (See [1] and related work in [2].) In this paper, we develop energy-efficient relaying protocols that exploit spatial diversity available at distributed antenna elements to combat fading due to multipath propagation, a particularly severe form of interference experienced in wireless networks.

To illustrate the main concepts, we consider the wireless network depicted in Fig. 1. At the physical layer, destination radios receive potentially useful signals from all transmitters that are active, and may combine multiple transmissions of the same signal to reduce variations in performance caused by signal fading, a technique referred to broadly as spatial diversity combining [3]. We refer to this form of spatial diversity as *distributed spatial diversity*, in contrast to the currently more conventional forms of spatial diversity [4], because the radios essentially share their antennas and other resources to create a "virtual array" through distributed transmission and signal processing.

After developing a mathematical model in Section 2 for the network in Fig. 1, we scratch the surface of the rich set of design issues and options that arise in the context of exploiting distributed spatial diversity for wireless networks. Section 3 casts the basic relaying protocols, referred to as direct and multihop transmission, respectively, into our framework, and explores a number of possibilities for diversity transmission and hybrid protocols, in terms of what signals the source and relay jointly transmit as well as how the relay and destination jointly process signals. Section 4 develops outage regions for the various transmission options, and Section 5 compares these regions and their corresponding outage probabilities. Performance comparisons in Section 5 suggest that our diversity transmission protocols are capable of overcoming the noisy channels between the distributed radio antennas to achieve diversity gain and outperform direct and multihop transmission in a variety of scenarios of interest.

2 System Model

Our network model consists a of collection of M radios that share L orthogonal channels. A transmission period consists of two consecutive blocks, and the channels are allocated to (up to) L radios during each transmission period. During the first block of a transmission period, a radio transmits on its assigned channel, and receives on a separate channel. Radios might choose the strongest channel or might select a channel at random for reception. Such options complicate the story for $L > 2$, so we consider the case of $L = 2$ for simplicity of exposition. Depending upon the strength of the signal received on the selected channel, the radio decides between resending its own transmission (or, more generally, additional parity bits from a more efficient code, *e.g.*, rate-compatible punctured codes) in the next block, or relaying the other radio's transmission in the next block. As a result, equal bandwidth and power allocations seem to be a natural choice.

In our model for the wireless network depicted in Fig. 1, narrowband transmissions suffer the effects of path loss and flat fading as arise in *e.g.*, slow-frequency-hop networks. Our analysis focuses on the case of slow fading to isolate the benefits of spatial diversity alone; however, we emphasize at the outset that our results extend naturally to the kinds of highly mobile scenarios in which faster fading is encountered.

Our baseband-equivalent discrete-time channel model for the network consists of two sub-channels, orthogonal in, *e.g.*, adjacent frequencies. This decomposition is necessary because practical limitations in radio implementation prevent the relays from simultaneously transmitting and receiving on the same channel. Thus, radio 1 transmits on channel 1 and receives on channel 2 to potentially relay the signal transmitted by radio 2, and vice versa. The received signals at all four radios are modeled by

$$\mathsf{y}_{j,i}[n] = \mathsf{a}_{i,j}\, \mathsf{x}_i[n] + \mathsf{z}_{j,i}, \quad i = 1, 2, \quad j = 1, 2, 3, 4, \quad j \neq i. \tag{1}$$

Here $\mathsf{a}_{i,j}$ captures the effects of path loss and static fading on transmissions from radio i to radio j, x_i is the transmitted signal of radio i having average energy $\mathcal{E}_i$, and $\mathsf{z}_{j,i}[n]$ models additive receiver noise and other forms of interference at receiver j in channel i.

Statistically, we model the fading coefficients $\mathsf{a}_{i,j}$ as zero-mean, mutually independent complex random variables with variances $\sigma^2_{\mathsf{a}_{i,j}}$, and we model the additive noises $\mathsf{z}_{j,i}[n]$ as zero-mean, mutually independent, white complex jointly Gaussian sequences with variance $\mathcal{N}_j$. The network geometry is assumed unknown or too dynamic to track, so the channels may be well modeled as having i.i.d. signal-to-noise ratios (SNRs) under an appropriate distribution. We denote the SNR in each received signal as $\gamma_{i,j} \triangleq |\mathsf{a}_{i,j}|^2\, \mathcal{E}_i/\mathcal{N}_j$. For example, under the Rayleigh fading model, the SNRs are independent exponential random variables.

As we develop our transmission protocol in Section 3, it will be convenient to consider successive pairs of data blocks from the channel model in (1). Specifically, for blocklength N, we collect the appropriate samples into the vectors

$$\begin{aligned} \mathbf{x}_i[k] &= \left[x_i[kN] \quad x_i[kN+1] \quad \cdots \quad x_i[kN+(N-1)]\right]^{\mathrm{T}}, \\ \mathbf{y}_{j,i}[k] &= \left[y_{j,i}[kN] \quad y_{j,i}[kN+1] \quad \cdots \quad y_{j,i}[kN+(N-1)]\right]^{\mathrm{T}}. \\ \mathbf{z}_{j,i}[k] &= \left[z_{j,i}[kN] \quad z_{j,i}[kN+1] \quad \cdots \quad z_{j,i}[kN+(N-1)]\right]^{\mathrm{T}}. \end{aligned} \tag{2}$$

3 Transmission Protocol

Throughout this paper, we focus on relatively simple protocols that operate on two consecutive blocks indexed by $2k$ (even) and $2k+1$ (odd). At a high level, our protocols involve the following steps:

- **Even Blocks:** Radios $i = 1, 2$ encode new information into blocks $\mathbf{x}_i[2k]$, respectively. Radios $j = 1, 2, 3, 4$ receive signals $\mathbf{y}_{j,i}[2k]$, $i = 1, 2$. Radios 3 and 4 defer their processing until the end of block $2k + 1$. The transmitting radios process their respective received signals and decide whether they will cooperate in the next block, and if so, how.

- **Odd Blocks:** Radios $i = 1, 2$ encode either their own or their partner's data into blocks $\mathbf{x}_i[2k+1]$, respectively. Radio 3 receives signals $\mathbf{y}_{3,i}[2k+1]$, $i = 1, 2$, and jointly processes these signals with $\mathbf{y}_{3,i}[2k]$, $i = 1, 2$, received in the previous block. Radio 4 operates on its respective received signals in similar fashion.

Among many possible coordination strategies, we consider a simple protocol in which the two cooperating radios accurately estimate the SNR $\gamma_{2,1} = \gamma_{1,2}$ between them and use this estimate to select a suitable cooperative action. Such a protocol allows the radios to retransmit (in the form of repetition codes or more sophisticated single-user coding schemes such as rate-compatible punctured codes) their own information when $\gamma_{1,2}$ is too small to justify cooperation, and to transmit each other's information (again, in the form of repetition codes or more powerful joint encoding schemes) when $\gamma_{1,2}$ is large enough to justify cooperation and, in particular, to provide spatial diversity benefit to the transmitting radios. In any case, a destination radio can appropriately combine its received signals by exploiting control information in the protocol headers, *e.g.*, a field indicating the cooperative action taken by the transmitting radios. While the nature and amount of this control information, as well as the accuracy and consistency of the SNR estimates at the two cooperating radios, are important practical considerations, a detailed study of their impact on system performance is beyond the scope of this paper.

Assuming both radios estimate the realized value $\gamma_{1,2}$ perfectly, they will choose identical cooperation strategies due to the statistical symmetry of the channels implied by our model. We examine three options for the transmissions in (odd) block $2k + 1$:

- **Direct Transmission:** The coordinating radios ignore each other's transmissions and re-transmit their information from block $2k$.

- **Multihop Transmission:** The coordinating radios fully decode and retransmit each other's information, and the destinations ignore the initial transmissions from block $2k$, processing only the relayed transmissions in block $2k + 1$.

- **Diversity Transmission:** The coordinating radios assist one another by transmitting each other' s information in block $2k+1$. Reasonable strategies explored in our previous work include decoding and forwarding as well as amplifying and forwarding.
  - **Decode and Forward:** If $\gamma_{1,2}$ is reasonably large, it is advantageous for the relay to decode the transmissions to suppress noise on the channel between the source and relay, and re-encode the signal, potentially with a different codebook, for transmission between the relay and destination.
  - **Amplify and Forward:** For situations in which $\gamma_{1,2}$ is small, a linear relay that simply amplifies its received signal can be shown to be more effective than decoding and forwarding.

To summarize, each protocol consists of a source codebook, a relay processing/coding scheme, and a destination decoder.

Our combined analysis and empirical studies suggest that we may employ a pair of threshold tests on the SNR between the cooperating radios to choose the strategy with best expected performance, as measure by, *e.g.*, the lowest (conditional) outage probability. To develop this result, we examine the outage regions associated with each case in Section 4, and compare conditional outage probabilities in Section 5. We stress at the outset that none of the protocols we propose are necessarily optimal, but they represent reasonable protocols whose performance we can evaluate and begin to optimize. Improving upon these protocols, and developing others, is the subject of on-going work.

4 Outage Regions

Generalizing upon our previous results for uncoded, *i.e.*, $N = 1$, single-user systems with a probability of bit-error performance measure [5], we characterize the performance of the various cooperation strategies in terms of outage regions and outage probabilities. Outage regions specified in terms of the SNR random variables $\gamma_{i,j}$ have convenient interpretations in both coded and uncoded settings, but we will develop our results from a coded perspective and determine events in which the realized capacity of the channel falls below a target transmission rate. We convert this event into an event defined in terms of the SNRs in the channel.

Since the capacity is a function of the SNR random variables of the channel, it too is a random variable. The event that this capacity random variable falls below some fixed rate R is referred to as an *outage event* (or *outage region* in terms of the SNR random variables), because reliable communication is not possible inside this region. The probability of an outage event is referred to as the *outage probability* of the channel,

$$P_{\text{out}}(R) = \Pr\left[C < R\right]. \tag{3}$$

We stress that outage regions are independent of the distribution of the underlying random variables, while outage probabilities are intimately tied to them. For example, if the outage region of a channel at a particular rate is a strict subset of the outage region of another channel at that rate, then the first channel has smaller outage probability regardless of the probability distribution on the channel parameters. Furthermore, as we will see, several of our cooperation strategies appear to have similar outage probabilities, but the structure of their outage regions is sufficiently different that we might prefer one over the other in various regimes. As a result, both outage regions and outage probabilities will be useful for characterizing our transmission protocols.

We consider transmission from radio s to radio d, with the potential of relaying the transmissions of radio r (and having radio r relay the transmissions of radio s). We parameterize the results in this form for compactness, but note that they can immediately be interpreted from the perspective of radio 1 by setting $s = 1$, $r = 2$, and $d = 3$; similarly, to interpret the results from the perspective of radio 2, we set $s = 2$, $r = 1$, and $d = 4$.

In the following sections, we determine the outage regions for direct transmission, multihop transmission, diversity transmission with amplifying and forwarding, and diversity transmission with decoding and forwarding.

4.1 Direct Transmission

Direct transmission in our setting corresponds to a point-to-point communications channel, to which we may readily apply classic information theoretic arguments [1]. Specifically, the capacity between the source and destination radios using repetition coding satisfies

$$C_{\mathrm{SH}} = C_{s,d} = \frac{1}{2}\log\left(1 + 2\gamma_{s,d}\right), \tag{4}$$

with $\mathbf{x}_s[2k] = \mathbf{x}_s[2k+1]$ (repetition coding) distributed as i.i.d. zero-mean complex Gaussian random variables each with variance $\mathcal{E}_s$. We note that, while in principle more powerful forms of coding than repetition are possible across the two blocks $\mathbf{x}_s[2k]$ and $\mathbf{x}_s[2k+1]$, comparison to diversity transmission with amplifying and forwarding, inherently analogous to repetition, is most convenient in the repetition coded case.

Inspecting (4), we see that the outage event $C_{\mathrm{SH}} < R$ can be readily manipulated into an event defined in terms of the SNR random variable between the source and destination, *i.e.*,

$$\gamma_{s,d} < t/2, \tag{5}$$

for an appropriate SNR threshold t that increases with increasing R.

4.2 Multihop Transmission

Multihop transmission corresponds to direct transmission between the source and relay radios followed by direct transmission between the relay and destination radios. Thus we might expect the capacity of the cascade of the two channels to be the minimum of the capacities. Indeed, results on cascade channels [1] yield

$$C_{\mathrm{MH}} = \min\{C_{s,r},\ C_{r,d}\} = \frac{1}{2}\min\{\log\left(1 + \gamma_{s,r}\right),\ \log(1 + \gamma_{r,d})\}, \tag{6}$$

where, again, equality is achieved for complex Gaussian signals, and where we lose the factor of 2 present in (4) because the destination only processes the signal received from the relay and ignores the signal transmitted by the source.

Again, we may determine the outage region corresponding to (6), yielding

$$\gamma_{s,r} < t \quad \text{or} \quad \gamma_{r,d} < t\ . \tag{7}$$

As we will see in Section 4.4, this outage region is a strict superset of the outage region for diversity transmission with decoding and forwarding. This observation allows us to eliminate multihop transmission from our comparisons in Section 5.

4.3 Diversity Transmission with Amplifying and Forwarding

Under diversity transmission with amplifying and forwarding, the relay scales its received sequence by

$$\beta = \sqrt{\frac{\mathcal{E}_r}{|a_{s,r}|^2\,\mathcal{E}_s + \mathcal{N}_r}}$$

to satisfy its average power constraint. Note that we allow the amplifier gain to depend upon the fading coefficient $a_{s,r}$ between the source and relay, which the relay estimates to high accuracy. This transmission scheme can be viewed as repetition coding from two separate transmitters, except that the transmitters may have different power levels and the relay transmitter actually amplifies its receiver noise. Nevertheless, the channel can be viewed as a single-user Gaussian noise channel and has capacity

$$C_{DA} = \frac{1}{2}\log(1 + \gamma_{s,d} + f(\gamma_{s,r}, \gamma_{r,d}))\ , \tag{8}$$

for x_s zero-mean complex Gaussian with variance $\mathcal{E}_s$, where

$$f(x, y) = \left[x^{-1} + y^{-1} + (xy)^{-1}\right]^{-1},$$

analogous to a parallel combination of resistances with values x, y, and xy, respectively.

The outage region becomes the SNR event

$$\gamma_{s,d} + f(\gamma_{s,r}, \gamma_{r,d}) < t\ , \tag{9}$$

and an outage region "boundary" for each realization of $\gamma_{s,r}$ follows after careful manipulation of (9), obtaining

$$\gamma_{r,d} = \begin{cases} 0, & \gamma_{s,d} \geq t \\ \frac{1+\frac{1}{\gamma_{s,r}}}{\frac{1}{t-\gamma_{s,d}} - \frac{1}{\gamma_{s,r}}}, & t - \gamma_{s,r} < \gamma_{s,d} < t \\ +\infty, & \gamma_{s,d} < t - \gamma_{s,r}. \end{cases} \ . \tag{10}$$

4.4 Diversity Transmission with Decoding and Forwarding

When the SNR between the source and relay is reasonably high, it is advantageous for the relay to decode the transmission and re-encode, potentially with a different codebook, for cooperative transmission to the destination. When the relay decodes successfully, the source and relay have the same information and can be viewed as antenna elements in a transmit diversity array that uses orthogonal channels (expands bandwidth). One can show that, under the constraint that the relay decodes perfectly, for $\mathbf{x}_r[2k+1] = \hat{\mathbf{x}}_s[2k]$, *i.e.*, the codebook is a repetition code, the capacity satisfies

$$C_{\mathrm{DD}} = \frac{1}{2}\min\{\log\left(1 + \gamma_{s,r}\right),\ \log\left(1 + \gamma_{s,d} + \gamma_{r,d}\right)\}\ . \tag{11}$$

Since we require the relay to decode perfectly, the outage regions correspond to the SNR event

$$\gamma_{s,r} < t \quad \text{or} \quad \gamma_{s,d} + \gamma_{r,d} < t, \tag{12}$$

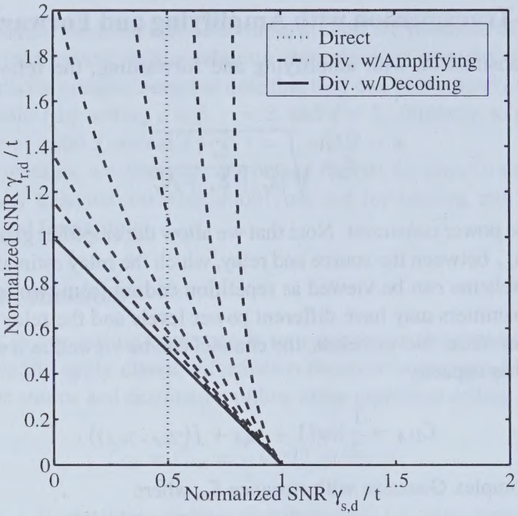


Figure 2: Outage region boundaries for repetition-code transmission protocols. The dotted line at $\gamma_{s,d}/t = 1/2$ corresponds to the outage region boundary for direct transmission. Successively lower dashed curves correspond to diversity transmission with amplifying and forwarding assuming received signal strength measurements of $\gamma_{s,r}/t = 1/4, 1/2, 1, 2, 4, 8$. The solid line corresponds to diversity transmission with decoding and forwarding for the case in which $\gamma_{s,r}/t > 1$; otherwise, the outage event for decoding is the entire plane. The outage regions lie below and/or to the left of the indicated lines, *i.e.*, all include the point $(0, 0)$.

which, as we observed in Section 4.2, is a subset of the outage region for multihop transmission. For each realization of $\gamma_{s,r}$, the boundary of the outage region (12) can be manipulated into the form

$$\gamma_{r,d} = \begin{cases} 0, & \gamma_{s,r} \geq t, \quad \gamma_{s,d} \geq t \\ t - \gamma_{s,d}, & \gamma_{s,r} \geq t, \quad \gamma_{s,d} < t \\ +\infty, & \gamma_{s,r} < t, \quad 0 \leq \gamma_{s,d} < \infty \end{cases} . \tag{13}$$

5 Protocol Design and Performance

Fig. 2 depicts the outage regions for direct and diversity transmission with repetition coding, and clearly indicates improved performance (successively smaller outage regions) for amplifying and forwarding with increasing $\gamma_{s,r}$, approaching the outage region of decoding and forwarding for $\gamma_{s,r}$ large relative to the threshold t. We will see that for a sufficiently large value of $\gamma_{s,r}$, the (conditional) outage probability for amplifying and forwarding will be smaller than that of direct transmission, even though the outage regions for direct and diversity transmission are not nested one way or the other. The same is true for comparisons between transmit antenna diversity systems (without beamforming) with one and two antennas. In practice, the dual antenna systems perform better than the single antenna systems.

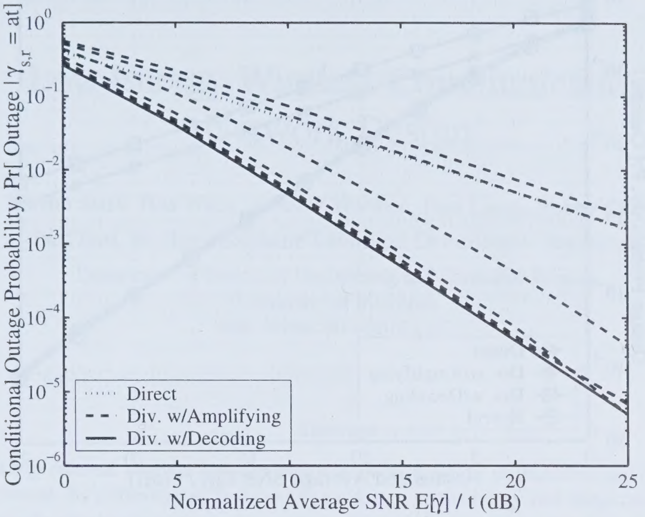


Figure 3: Conditional outage probabilities under exponential statistics. The dotted curve corresponds to direct transmission, while the solid curve corresponds to diversity transmission with decoding and forwarding for $\gamma_{s,r}/t \geq 1$. (For $\gamma_{s,r}/t < 1$, the conditional outage probability of diversity transmission with decoding and forwarding is 1.) Successively lower dashed curves correspond to diversity transmission with amplifying and forwarding for $\gamma_{s,r}/t = 1/4, 1/2, 1, 2, 4, 8$.

Indeed, conditional outage probability calculations appear to confirm this hypothesis. For example, Fig. 3 shows outage probability calculations for the various transmission strategies, assuming the SNRs are i.i.d. exponential random variables with mean $\mathrm{E}[\gamma]$. Moreover, the results in Fig. 3 suggest thresholds for selecting among the various transmission strategies. Specifically, for $\gamma_{s,r}/t < 1/2$, these results suggest that direct transmission is preferable. For $1/2 \leq \gamma_{s,r}/t < 1$, bearing in mind that the conditional outage probability for diversity transmission with decoding and forwarding is 1 given $\gamma_{s,r}/t < 1$, we should employ diversity transmission with amplifying and forwarding. Finally, for $\gamma_{s,r}/t \geq 1$, the decoding relay offers uniformly lower conditional outage probability than direct transmission or diversity transmission with amplifying and forwarding.

Fig. 4 shows empirical outage probabilities for our various protocols for the case in which all the SNRs in the channel are i.i.d. exponential random variables with mean $\mathrm{E}[\gamma]$. These results suggest that energy savings on the order of 10 dB at $P_{\text{out}} = 10^{-3}$ can be obtained using protocols developed in this paper, because these protocols efficiently create multiple, independently-faded transmissions at separate radios that can be effectively combined at the destination receiver to achieve diversity gains. We also note that a protocol that always employs diversity transmission with amplifying and forwarding performs almost as well as our optimized hybrid protocol.

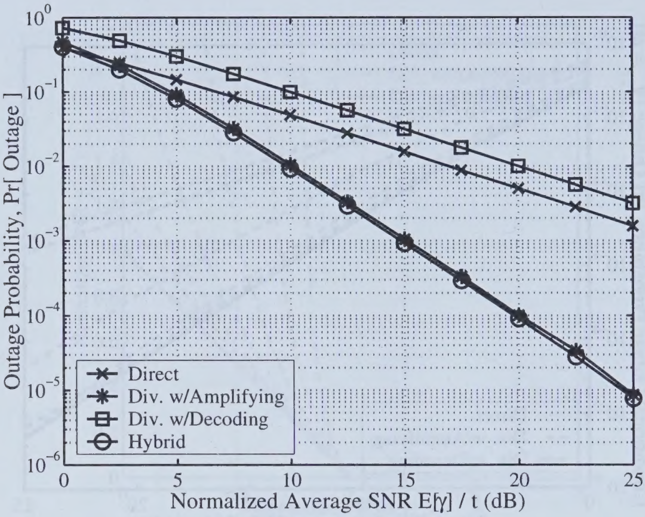


Figure 4: Outage probabilities under exponential statistics.

References

[1] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, John Wiley & Sons, Inc., New York, 1991.

[2] Andrew Sendonaris, Elza Erkip, and Behnaam Aazhang, "Increasing uplink capacity via user cooperation diversity," in *Proc. IEEE Internation Symposium on Information Theory (ISIT)*, Cambridge, MA, Aug. 1998.

[3] John G. Proakis, *Digital Communications*, McGraw-Hill, Inc., New York, Third edition, 1995.

[4] Aradhana Narula, Mitchell D. Trott, and Gregory W. Wornell, "Performance limits of coded diversity methods for transmitter antenna arrays," *IEEE Trans. Inform. Theory*, vol. 45, no. 7, pp. 2418–2433, Nov. 1999.

[5] J. Nicholas Laneman and Gregory W. Wornell, "Energy-efficient antenna sharing and relaying for wireless networks," in *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*, Chicago, IL, Sept. 2000.

Low Energy Wireless Communication Network Design*

Wayne Stark, Hua Wang, Andrew Worthen, Paul Liang, Robby Gupta,
Jack East, Al Hero, Stéphane Lafortune, Demosthenis Teneketzis
Department of Electrical Engineering and Computer Science
University of Michigan
Ann Arbor, MI 48109-2122

Abstract

Energy efficient wireless communication network design is an important and challenging problem. Its difficulty lies in developing a design methodology that integrates various aspects of a wireless communication network from the device layer, such as antenna and amplifier design, all the way to the network and application layers. We present a generic simulation-based design methodology that achieves global optimization for such networks and that is suitable for many kinds of mobile systems. We point out some important connections between the different layers in the design procedure. Some preliminary results of the application of our design methodology to a situational awareness mobile wireless network are given. These results illustrate the trade-off between performance and energy consumption in the operation of these networks.

1 Introduction

Energy efficient wireless communication network design is an important and challenging problem. It is important because mobile units operate on batteries with limited energy supply. It is challenging because there are many different issues that must be dealt with when designing a low energy wireless communication system (such as amplifier design, coding, and modulation design), and these issues are coupled with one another. Furthermore, the design and operation of each component of a wireless communication system present tradeoffs between performance and energy consumption. Therefore, the challenge is to exploit the coupling among the various components of a wireless communication system, and understand tradeoffs between performance and energy consumption in each individual component, in order to come up with an overall integrated system design that has optimal performance and achieves low energy (power). The key observation is that constraining the energy of a node imposes a coupling among the design layers that can not be ignored in performing system optimization. In addition, the coupling between layers requires simulation in order to accurately determine the

---

*This research is supported by the Department of Defense & Engineering (DDR&E) Multidisciplinary University Research Initiative (MURI) on "Low-Energy Electronic Design for Mobile Platforms" and managed by the Army Research Office (ARO) under grant DAAH04-96-1-0377.

performance. The purpose of this paper is to present a methodology for the design, simulation and optimization of wireless communication networks for maximum performance with an energy constraint.

Before we proceed, we illustrate, through simple examples, a couple of issues that need to be addressed. To highlight the tradeoffs between performance and energy consumption at individual components, consider the design and operation of an amplifier. The amplifier boosts the power of the desired signal so that the antenna can radiate sufficient power for reliable communications. However, typical power amplifiers have maximum efficiency in converting DC power into RF power when the amplifier is driven into saturation. In this region of operation, the amplifier voltage transfer function is nonlinear. Because of this nonlinearity, the amplifier generates unwanted signals (so called intermodulation products) in the band of the desired signal and in adjacent bands. When the amplifier drive level is reduced significantly (large backoff) the amplifier voltage transfer characteristic becomes approximately linear. In this case it does not generate intermodulation products. However, with large backoff the amplifier is not able to efficiently convert DC power into RF power. Thus, there is considerable wasting of power at low drive levels, but at high drive levels more interfering signals are generated.

To highlight the coupling among the design of individual components of a wireless system, consider packet routing in a wireless network that contains no base stations (i.e. an ad hoc network). For simplicity consider a network with nodes A, B, and C shown in Figure 1.1. If Node A wants to transmit a message to Node C, it has two options. Transmit with power sufficient to reach Node C in a single transmission, or transmit first from A to B with smaller power, and then from B to C. Since the received signal power typically decays with distance as d^4, there is significantly smaller power loss due to propagation in the second option because $d_{AC}^4 > d_{AB}^4 + d_{BC}^4$. However, even though Node A transmits with smaller output power, it does not necessarily proportionally decreases the amount of power actually consumed because of the amplifier's effect discussed above. Furthermore, besides the energy required for packet transmission, there are energy requirements for packet reception and information decoding. The probability of packet error reception that is achieved depends on the energy allocated to the receiver. Consequently, there is a coupling among amplifier design, coding, and modulation design, and decoding design as well as routing protocol.

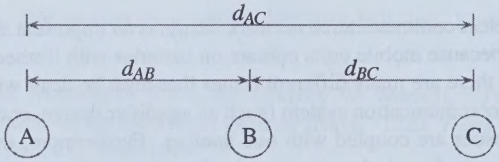


Figure 1.1: Nodes in a network

Finally, to highlight the coupling and tradeoff between energy consumption and system wide performance, we consider the situational awareness problem in a mobile wireless network. In this problem, the objective of each node is to be aware of the position of every other node. If energy consumption is ignored, and the overall performance metric is an average (over all mobiles) mean squared position error, this error is minimized when all nodes continuously communicate their positions with one another. Such a strategy requires significant energy. If, on the other hand, the objective is to minimize a weighted sum of the expected energy consumed and the average (over all mobiles) mean squared position error, then the nodes have to jointly decide when to communicate and whom to communicate to. Quantifying the trade-off between information flow and performance is a nontrivial task, and it is one of the main goals of our research effort. Such a quantification requires the understanding of all the issues highlighted by all the previous examples.

The remainder of the paper is organized as follows. In Section 2 we present a methodology for system design incorporating the effect of the amplifiers, propagation, and network layer protocols. This methodology is fairly general and can be applied to many different applications besides the situational awareness system we consider. In Section 3 we describe the model for the amplifier, propagation, coding, modulation and network protocol. In Section 4 we present the results we obtained for the situational awareness application. Section 5 concludes the paper.

2 System Design Methodology

We consider a wireless network consisting of interacting mobiles or nodes. The mobiles need to communicate in order to take some action or to share information, such as their respective positions. In this section, we present our simulation-based design methodology that simultaneously optimizes a number of different parameters including, amplifier drive level, quantization levels at the receiver for demodulation and decoding, and network protocol parameters. The overall goal is to characterize the performance as a function of energy. In order to achieve the desired design and optimization we divide the problem into interacting design layers as shown in Figure 2.1. Each layer has a local optimization but also interacts with layers above and below in a well defined manner. In addition, there is a global optimization that encompasses the different layers. The lowest layer is the device layer. This layer contains the antenna model and the amplifier model. The next layer is the processing layer. This layer contains a model for the physical layer communication system including the coding algorithm and the demodulation. The top layer is the network layer which accounts for the routing protocol and contains a model of the mobile locations. We also include in the network layer the overall energy consumption and performance metric.

From an overall perspective the design methodology is described as follows:

- Divide the system into hierarchical layers.
- Identify the direct interactions between layers. In our case, for example, the amplifier drive affects the packet error probability through the transmitted power and the energy consumed through the DC power. Indirect interactions will "trickle through" the model.
- Optimize each layer as a function of its interactions with the other layers. In our example, we select the bit allocations for minimum packet error rate subject to particular energy consumption constraints.
- Using simplified models of the layers based on the individual optimizations, optimize the complete system. These simplified models could be based on table lookups. This step captures complex interactions between the layers, with the multi-input, multi-output character of each layer capturing the complex trade-off in each element's design.
- The optimizations at the different layers may have to be simulation-based in the absence of complete analytical expressions for performance or of necessary convexity assumptions for the available analytical expressions.

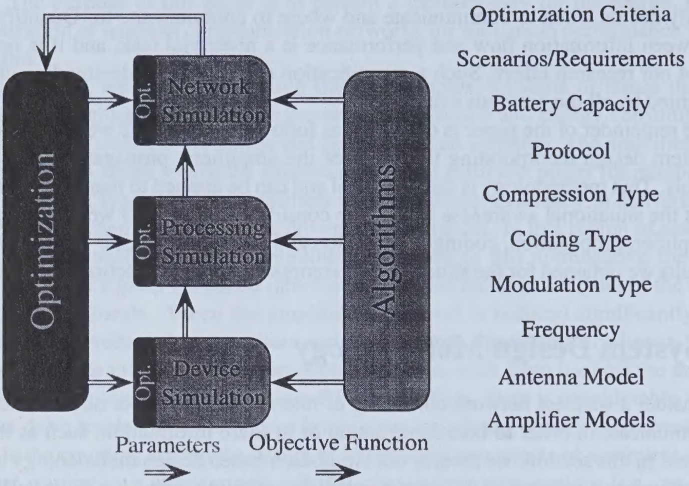

Optimization Criteria
Scenarios/Requirements
Battery Capacity
Protocol
Compression Type
Coding Type
Modulation Type
Frequency
Antenna Model
Amplifier Models
Optimization
Opt.
Network Simulation
Opt.
Processing Simulation
Opt.
Device Simulation
Algorithms
Parameters
Objective Function

Figure 2.1: Layered Design/Optimization

The general method we use to optimize this system is as follows. First a model for the amplifier characteristics is determined as a function of the amplifier drive level and bias. This determines the nonlinear characteristics of the amplifier. It also determines the energy efficiency (DC energy to RF energy) of the amplifier. In addition it determines the amount of intermodulation products. The intermodulation products are generated by the nonlinearity of the amplifier. Some of the intermodulation products fall into the band of the desired signal and create in-band interference while some of the intermodulation products fall into adjacent bands and create interference to users in adjacent bands. In Figure 2.2 we show an amplifier nonlinearity and power efficiency characteristics.

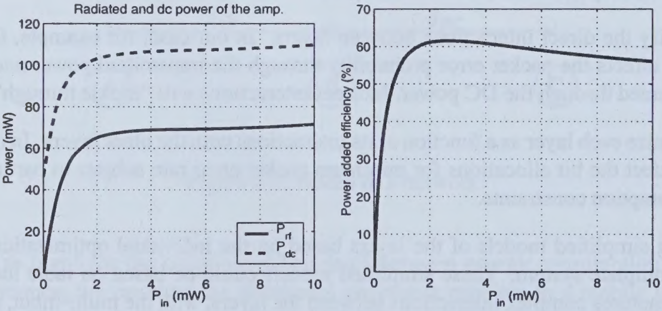


Figure 2.2: Amplifier Characteristics

In the methodology shown in Figure 2.1 the inputs to the device layer are the parameters of the amplifier and the outputs of the device layer are the desired power out, the DC power consumed, the in-band intermodulation power and the adjacent band intermodulation power. These parameters are used by the processing layer to determine a packet error probability.

At the processing layer we need to account for the physical layer of a communication system. This includes the coding algorithm and the decoding algorithm. We must also incorporate the effects of the nonlinear amplifier as determined by the device layer. The coding and modulation block diagram is shown in Figure 2.3.

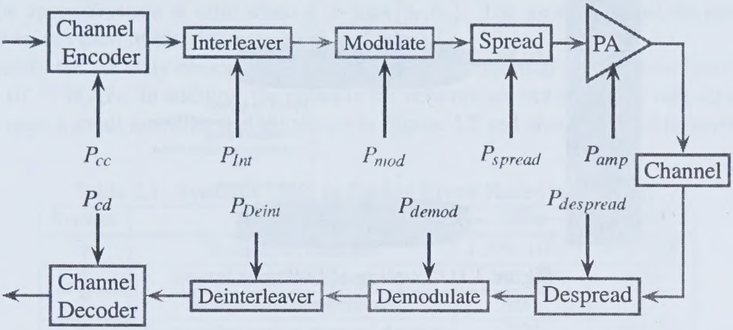


Figure 2.3: Processing Layer Block Diagram

In the block diagram we note that each block consumes energy. However we focus our attention on the energy used by the demodulator and the decoder as well as the amplifier power consumption.

At the network layer we consider the routing protocol, information distribution issues, communiçation environment, mobility modeling and overall performance measurement. We need to incorporate the effects of the processing layer to account for the decoding capability and energy consumption, which in turn affects the network performance.

3 Simulation-Based Optimization – System Overview

The performance evaluation including the effects of the amplifier through the network protocols could be extremely time consuming. As such an efficient technique is required to evaluate performance incorporating the effects of all the layers. We demonstrate our design and performance evaluation methodology by applying it to the optimization of a particular network application, namely, the situational awareness problem in mobile networks. The actual parameters passed between the different layers for this problem are shown in Figure 3.1. The division of layers is for conceptual purposes only. Besides each local parameterized optimization of the processing and device layer, a global optimization is run over the integrated system as can be seen on the left side of Figure 3.1. The interaction of different layers is through tabular forms, as illustrated in Figure 3.2. We proceed to describe in detail the models used in each layer in a top-down manner.

3.1 Network Layer

We have simulated a situational awareness scenario, where a network of nine nodes, initially deployed in a drop zone, travel toward the same destination. While traveling, each node informs the other nodes of its own position and also estimates the other nodes' positions periodically.

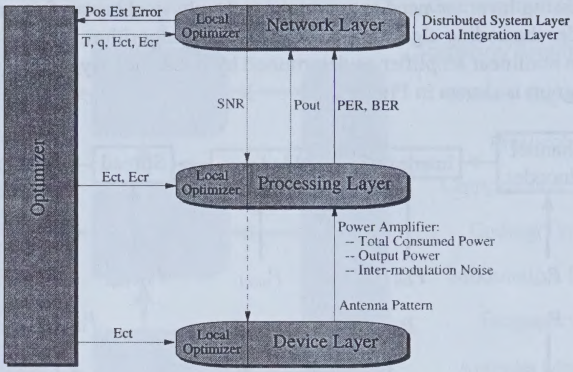

Optimizer
Pos Est Error
T, q, Ect, Ecr
Local Optimizer
Network Layer
Distributed System Layer
Local Integration Layer
SNR
Pout
PER, BER
Ect, Ecr
Local Optimizer
Processing Layer
Power Amplifier:
-- Total Consumed Power
-- Output Power
-- Inter-modulation Noise
Antenna Pattern
Ect
Local Optimizer
Device Layer

Figure 3.1: Coupling of Different Layers

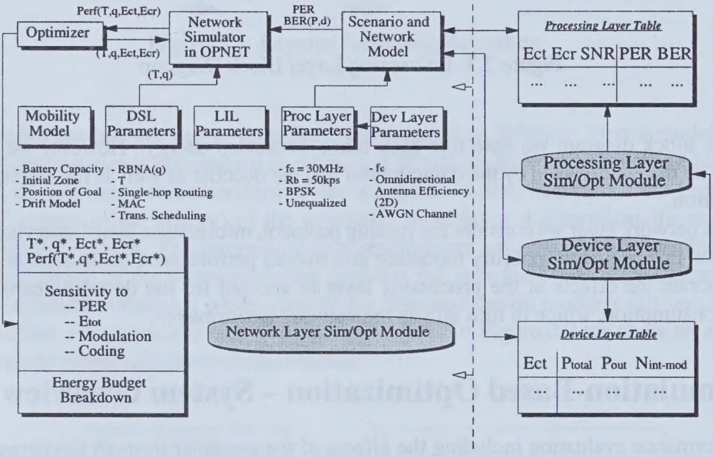

Perf(T,q,Ect,Ecr)
Optimizer
(T,q,Ect,Ecr)
Network Simulator in OPNET
(T,q)
PER
BER(P,d)
Scenario and Network Model
Processing Layer Table
Ect Ecr SNR PER BER
Mobility Model
DSL Parameters
LIL Parameters
Proc Layer Parameters
Dev Layer Parameters
- Battery Capacity
- Initial Zone
- Position of Goal
- Drift Model
- RBNA(q)
- T
- Single-hop Routing
- MAC
- Trans. Scheduling
- fc = 30MHz
- Rb = 50kps
- BPSK
- Unequalized
- fc
- Omnidirectional Antenna Efficiency (2D)
- AWGN Channel
Processing Layer Sim/Opt Module
Device Layer Sim/Opt Module
T*, q*, Ect*, Ecr*
Perf(T*,q*,Ect*,Ecr*)
Sensitivity to
-- PER
-- Etot
-- Modulation
-- Coding
Energy Budget Breakdown
Network Layer Sim/Opt Module
Device Layer Table
Ect Ptotal Pout Nint-mod

Figure 3.2: Overview of Simulation-Based Optimization

Each node carries a battery with a certain capacity and a dipole antenna driven by an amplifier. All antennae are of the same height. A node can not transmit or receive any packet if its battery has been exhausted. Our objective is to optimize a set of parameters for a given communication strategy in order to obtain the performance versus energy tradeoff, and then to compare the performance of different communication strategies.

Each node moves with speed $1m/s$ and changes its location every $1s$. At every second, the final location of each node is its previous location plus a $1m$ movement toward the final destination plus a random component.

The transmitted signal from each node experiences fading. In the results to follow, we assume a two-path propagation model. It consists of a direct path and a multipath due to reflection off the ground with 180 degree phase change at the reflection point from the transmitter to the receiver. The cumulative effect of both paths gives us a relation between received power and transmitted power, which is given by

$$P_r = P_t 4 \left(\frac{\lambda_c}{4\pi d} \right)^2 G_t G_r \sin^2 \left(\frac{2\pi h_t h_r}{\lambda_c d} \right) \approx P_t G_t G_r \frac{h_t^2 h_r^2}{d^4} \tag{3.1}$$

where the approximation is valid when $d \gg \max\{h_t, h_r\}$. The meaning of all the parameters and their values used in simulation is given in Table 3.1.

Currently we are only considering thermal noise. The thermal noise power level is $N_0 = kT = 4 \cdot 10^{-21}\ W/Hz$. In addition, the nodes in the network are not separated very far apart, so we only need a small amplifier model (shown in Figure 2.2 and also described in Section 3.3).

Table 3.1: **Symbols Used in Packet Error Rate Calculation**

Symbol	Meaning	Value	Unit
k	Boltzmann constant	$1.379 \cdot 10^{-23}$	$W/Hz/K$
C	propagation speed	$3.0 \cdot 10^8$	m/s
T	background temperature	290	K
N_0	single-sided power spectral density	-204	dBW/Hz
f_c	carrier frequency	30	MHz
λ_c	carrier wavelength	10	m
h_t	height of transmitter antenna	1	m
h_r	height of receiver antenna	1	m
G_t	gain of transmitter antenna	1	*none*
G_r	gain of receiver antenna	1	*none*
d	propagation distance	$(0, 9000)$	m
P_t	transmitted power	$[10^{-7}, 1]$	W
P_r	received power	Eqn (3.1)	W

3.2 Processing Layer

Coding and modulation schemes have been designed and locally optimized at the processing layer in order to mitigate small-scale fading and thermal noise. For the purpose of coupling the processing layer with the network layer, the optimal performance of coding and modulation schemes have been parameterized with respect to the transmitter energy constraint and to the receiver energy constraint for demodulation and decoding of each packet. In this way, the overall network simulation can choose the appropriate transmit energy and receive energy for each packet.

We have evaluated various coding and modulation schemes for performance versus energy consumption [2]. In particular, convolutional codes and turbo codes were designed and their performance and energy consumption were derived for use by the network layer simulation. The packet length for situational awareness information is 224(252) bits, and these information bits are encoded using a $(2,1,7)$ convolutional code (or a turbo code) to form 460(512) channel bits. These channel bits are modulated by BPSK and the modulated signal is shaped by a square root raised cosine filter with rolloff rate 0.3, which is implemented using 19 taps with infinite quantization. The symbol duration is 20 μs and the waveform of each symbol is sent over a frequency-hopped spread spectrum communication system with 23(32) bits per hop, i.e., 20(16) hops per packet. At the receiver, the received signal and the additive white Gaussian noise are sampled at four times the transmission rate ($50kb/s$) before being passed through a matched filter with a tap delay line model with 19 taps and N_D bits of quantization for the coefficients and input samples. The decoder is a Viterbi algorithm with N_E bits of quantization. At the processing layer, we optimize the number of bits of quantization for the matched filter and decoder. We have observed that the overall network performance is actually worse if nodes use turbo decoder than if nodes use convolutional decoder for the situational awareness scenario we investigated, which has relatively small packet lengths.

3.3 Device Layer

An omni-directional dipole antenna together with a small amplifier were constructed at the device layer for transmission. The amplifier is usually the component that consumes the most energy in a wireless device. In our simulation we have a relatively small power amplifier model whose characteristics are tabulated and used by the processing layer. As already pointed out in Section 2, when the input power to the amplifier is relatively low, we have input-output linearity; however the amplifier operates at a very low efficiency, i.e., the ratio of useful output energy to consumed energy is low. If we drive the amplifier harder, with higher input power, the amplifier operates at higher efficiency, however with large input-output non-linearity. This non-linearity causes inter-modulation interference with in-band and out-band signals, which adversely affects the performance of the processing layer, and in turn affects the design of the network layer protocols.

When we design efficient low-energy networks, what we really want is to minimize the energy drained out of the battery, since this energy is what ultimately affects the battery life. The amplifier efficiency directly affects the consumed energy. In order to have any power out of the amplifier, there is a minimum amount of energy that has to be used. In current literature on energy efficiency of routing protocols, the transmitted energy (power) is usually chosen as the routing metric. But this choice does not reflect the actual consumed energy. In Section 4, we show that only considering transmitted energy in wireless routing protocol design is not good enough, together with the fact that a node needs to consume a certain amount of energy to run the routing protocol, the network performance is worse in multi-hop routing than in one-hop transmission for the case of a small drop zone.

3.4 Interactions Among System Layers

We now explain the interactions between the layers as shown in Figure 3.1 and Figure 3.2 for our simulation-based optimization on the one-hop transmission case. Our optimization and simulation iterate over step 1 to 4 to find the global optimum solution.

- Step 1: Optimization program determines the parameters E_{ct}, E_{cr}, T, q.

- Step 2: Device layer determines P_{out} (amplifier output power).

$$P_{out} = f_1(E_{ct}) \tag{3.2}$$

  where E_{ct} is transmitter energy constraint per packet.

- Step 3: Network layer determines SNR (signal-to-noise ratio).

$$P_r = P_{out} 4 \left(\frac{\lambda_c}{4\pi d} \right)^2 G_t G_r \eta_t \eta_r \sin^2 \left(\frac{2\pi h_t h_r}{\lambda_c d} \right) \tag{3.3}$$

$$SNR = \frac{P_r \cdot T_s}{N_0} \tag{3.4}$$

where P_r is received power out of receiver antenna, T_s is channel symbol duration, and N_0 is thermal noise power spectral density.

- Step 4: Processing layer determines *PER* (packet error rate), which is computed by calling the function f_4^* (this function uses the table generated by the processing layer).

$$PER = f_4^*(E_{ct}, E_{cr}, SNR) \tag{3.5}$$

where E_{cr} is receiver energy constraint per packet.

It is critical to understand that the simulation effort has been carefully divided into a device layer simulation, processing layer simulation, and network layer simulation. Furthermore, the interaction between layers has been captured.

4 Numerical Result

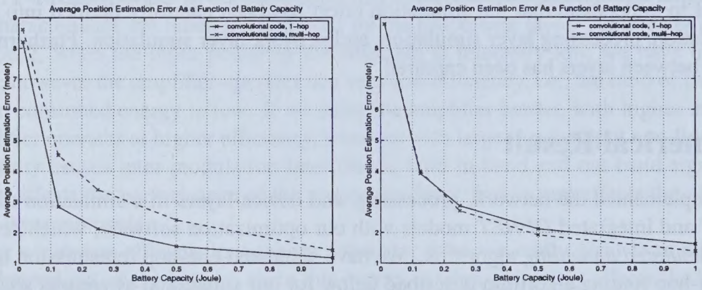
We have implemented the network, processing, and device layers in a simulation tool based on *OPNET* and integrated *OPNET* models with our optimization software, which consists of a type of *simulated annealing* algorithm. We have simulated one-hop transmission broadcast and a multi-hop routing algorithm described below for our situational awareness scenario for different channel coding methods.

For the one-hop transmission strategy, each node simply transmits a packet without any acknowledgment from the receivers. Since each node has an omni-directional antenna, the packet is broadcasted in all directions. The node which receives the packet does not forward any packet. However, the transmitting node retransmits the transmitted packet with a certain probability. The optimization is done over the parameter set (T, q, E_{ct}, E_{cr}), where T is the period of transmission, q is the retransmission probability, E_{ct} is transmitter energy constraint per packet, E_{cr} is receiver energy constraint per packet. The performance measure is mean squared estimation error for the positions of the mobiles.

We have implemented a link state routing algorithm that resembles the OSPF protocol [3, 4] used in the Internet. The link cost for the routing algorithm is defined to be d^4, where d is the distance between the transmitter and the receiver. The algorithm finds the shortest path tree from the sender to all other nodes. The link state routing algorithm requires flooding of the information of each link (depending on transmitted power) incident to a node. The flooding period T_p for the link information is one of the optimizing parameters. The situational awareness information is transmitted by each node using power cd^4 every T_a seconds, where T_a is the period between each transmission and c is an adjusting factor for transmitted power that we want to determine from our simulation-based optimization. In summary, for this multi-hop routing algorithm, the parameter set (T_p, T_a, c, E_{cr}) is used for the global optimization.

4.1 Performance Versus Energy Tradeoff

The performance of estimation error versus battery capacity is shown in Figure 4.1. As can be seen, for the same convolutional decoding method, the one-hop transmission method actually does better than the multi-hop routing algorithm when the nodes are initially dropped in a $1km \times 1km$ zone, which does not agree with our anticipation if we only consider transmitted energy. However, when the nodes are dropped in a $2km \times 2km$ zone, the multi-hop routing algorithm does better than the one-hop transmission method if given enough energy. This appears to be right if we take into account of the nonlinearity of the power amplifier and consumed energy for running routing protocol. For the one-hop transmission method, about 30% of the consumed energy is used for transmitting packets for both cases of drop zones, with the other 70% used for receiving packets. For the multi-hop routing, about 25% of consumed energy is used for transmitting packets for the case of $1km \times 1km$ drop zone, while about 35% of consumed energy is used for transmitting packets for the case of $2km \times 2km$ drop zone. The protocol uses about 40-70% of the consumed energy for transmitting and receiving packets, which are used only for updating routing information, in the scenario with multi-hop routing under different battery constraints.



(a) Drop Zone $1km \times 1km$ (b) Drop Zone $2km \times 2km$

Figure 4.1: Performance Comparison of Different Algorithms with Different Drop Zones

5 Conclusions

We have demonstrated the power of our simulation-based optimization and design methodology for wireless mobile networks by applying it to a situational awareness problem. We pointed out the important interconnections between the different layers (device, processing, network) that arise in designing a complex mobile system. We showed how to account for these interconnections in order to obtain an integrated design that achieves optimal performance.

References

[1] Theodore Rappaport, Wireless Communications: Principles and Practice. Englewood Cliffs, NJ: Prentice Hall, 1996, pp69-122, 139-196.

[2] Andrew Worthen, Sangjin Hong, Riten Gupta, and Wayne Stark, "Performance Optimization of VLSI Transceivers for Low-Energy Communications Systems," *Proc. IEEE MILCOM*, vol. 2, pp1434 -1438, October 1999.

[3] John Moy, "OSPF Version 2," *Internet RFC 2328*, April 1998.

[4] John Moy, "Multicast Extensions to OSPF," *Internet RFC 1584*, March 1994.

A Comparison of Signal Design Methods for Correlated Waveform Multiple-Access Communications

Tommy Guess[†] and Mahesh K. Varanasi[‡]
[†] Department of Electrical Engineering, Univ. of Virginia, VA 22904-4743
[‡] Department of ECE, Univ. of Colorado, Boulder, CO 80309-0425

38th Annual Allerton Conference on Communication, Control, and Computing, October 4-6, 2000

Abstract

There have been several recent papers in the literature that deal with the design of signature waveforms for use by the transmitters for up-link, single-cell, multiple-access communications. In particular, we consider the approach introduced in [1] and [2] where the signature waveforms are specifically designed for the centralized multiuser receiver at the base so that each transmitter can be guaranteed a pre-assigned Quality-of-Service (QoS) requirement in terms of the received Signal-to-Interference Ratio (SIR). The resulting strategy is called Bandwidth Efficient Multiple Access (BEMA) [2]. When all users employ Pulse Amplitude Modulation (PAM) and a common signaling rate, the key question in BEMA is how the waveforms must be designed to occupy as little bandwidth as possible and still meet the QoS objectives. For a strict measure of bandwidth, and for a given set of received powers, this question was addressed by the authors for the Maximum Signal-to-Interference Ratio-Decision Feedback (MSIR-DF) receiver in [3]. A similar question was addressed in [4] where, for a sum constraint on the received powers, optimal signature signals and transmit powers were obtained for the *linear* MSIR receiver (without decision feedback). A somewhat different but related non-QoS approach proposes the design of signature signals that maximize the total capacity of the multiple-access channel under a spreading gain constraint (cf. [5]-[6]). This paper undertakes a comparison of the minimum bandwidth required to achieve the QoS requirements for the signals designed for the MSIR-DF receiver [3], the linear MSIR receiver [4], and for sum-capacity maximization as in [6]. We show that the bandwidth required for multiuser receivers with decision feedback can be significantly less than that required for linear receivers or for sum-capacity maximization.

I. THE CWMA SYSTEM MODEL

A Correlated-Waveform Multiple Access (CWMA) system is modeled as one where K users transmit PAM waveforms simultaneously and the received signal is a symbol-synchronous superposition of those waveforms and additive noise so that

$$y(t) = \sum_n \sum_{k=1}^{K} X_k(n) u_k(t - nT) + n(t), \tag{1}$$

where $u_k(t)$ and $\{X_k(n)\}_n$ are the signature waveform and the sequence of transmitted symbols of the k^{th} user, respectively, and $n(t)$ is an additive white Gaussian noise process with a two-sided power spectral density of $N_0/2$. The signature waveforms are normalized to have unit energy so that the received power of the k^{th} user in units of energy/symbol is given by $p_k = E[X_k(n)^2]$. Define the diagonal power matrix $\mathbf{P} = \text{diag}(p_1, \cdots, p_K)$. It is assumed that the signature waveforms satisfy the Multiuser Nyquist Condition (MNC) which we define as:

$$\int_{-\infty}^{\infty} u_i(t) u_j(t + nT) dt = R_{ij} \delta_n \, , \tag{2}$$

where δ_n is the Kronecker delta (equal to 0 when $n \neq 0$ and 1 when $n = 0$). Define the signal correlation matrix $\mathbf{R} = \{R_{ij}\}$ (note the diagonal elements are equal to unity).

This work was supported in part by NSF grant NCR-9706591.

An equivalent discrete-time model can be obtained by passing the received signal through a bank of filters matched to the users' signature waveforms and then sampling the outputs at integer multiples of T. The sufficient statistics thus obtained can be expressed as

$$\underline{Y}(n) = \mathbf{R}\underline{X}(n) + \underline{N}(n), \tag{3}$$

where $\underline{X}(n)$ is the vector of information symbols of the users at time n and $\{\underline{N}(n)\}$ is a sequence of independent, identically-distributed, zero-mean, Gaussian random vectors, each with covariance $\frac{N_0}{2}\mathbf{R}$. Note that the CWMA channel is memoryless as a consequence of the MNC constraint. We drop the time index n and work with the model $\underline{Y} = \mathbf{R}\underline{X} + \underline{N}$.

The MNC is trivially satisfied if the signature waveforms are time-limited to a symbol duration but for a finite value of the strict measure of bandwidth, these signals can not be finitely supported. It can be shown that the minimum (strict) bandwidth required to satisfy the MNC is equal to $M/(2T)$ where M is the rank of $\mathbf{R}$. Given $\mathbf{R}$, it is possible to construct the signature wavforms that satisfy the MNC as linear combinations of M orthonormal basis functions. These basis functions can be the M time-translates of a single sinc pulse of bandwidth $M/(2T)$ (translated by integer multiples of T/M), or M modulations of the sinc function of bandwidth $1/4T$ (modulated by frequencies that are positive and odd integer multiples of $1/4T$) [7]. With these choices for the basis functions, note that when $\mathbf{R}$ is equal to the identity matrix, the CWMA model reduces to TDMA or FDMA, respectively.

II. BANDWIDTH-EFFICIENT MULTIPLE ACCESS

In this section, we briefly summarize the signal design method for multiuser receivers with decision feedback as obtained by the authors in [3].

A decision feedback multiuser receiver detects (or decodes in a coded system) the users successivly and is shown in Figure 1. It is parameterized by a feedforward matrix $\mathbf{F}$ and a strictly lower-triangular feedback matrix $\mathbf{B}$. The first user is decoded based on the decision statistic $Z_1 = \underline{F}_1^T\underline{Y}$ where $\underline{F}_1^T$ is the first row of $\mathbf{F}$ (the superscript T denotes transpose of a matrix). The resulting symbol decision is denoted as $\hat{X}_1$. The k^{th} user is decoded based on the decision statistic $Z_k = \underline{F}_k^T\underline{Y} - \sum_{i=1}^{k-1} B_{ki}\hat{X}_i$. Such processing effectively provides for each

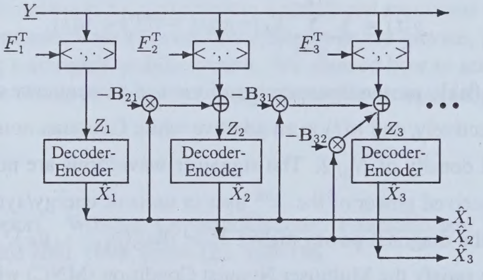


Fig. 1. The decision-feedback receiver.

user a single-input single-output channel. For a given order in which the users are processed, and under the assumption of perfect feedback, $\mathbf{F}$ and $\mathbf{B}$ are obtained to maximize every users' signal-to-interference ratios (SIR) in [8] [9]. This receiver will be refered to as the MSIR-DF receiver and is specified next.

Define the weighted correlation matrix $\mathbf{H} = \mathbf{P}^{1/2}\mathbf{R}\mathbf{P}^{1/2}$. Let $\mathbf{H}_{(k+1)}$ denote the principal submatrix of $\mathbf{H}$ formed by the indices k through K, and $\underline{H}_{(k)}$ the first column of $\mathbf{H}_{(k+1)}$ after striking out its first element.

The feedforward matrix for the MSIR-DF receiver is given by $\mathbf{F} = (\mathbf{L}^{\mathrm{T}})^{-1}\mathbf{P}^{1/2}$ where $\mathbf{H} + \frac{N_0}{2}\mathbf{I} = \mathbf{L}^{\mathrm{T}}\mathbf{L}$ is a Cholesky decomposition with $\mathbf{L}$ being lower triangular, and where $\mathbf{I}$ is the identity matrix. The feedback matrix is the strictly lower triangular part of the matrix product $\mathbf{FR}$. The SIR of the k^{th} user for the MSIR-DF receiver, γ_k, can be expressed as [7]

$$\gamma_k = (\tfrac{N_0}{2})^{-1}\left(p_k - \underline{H}_{(k)}{}^{\mathrm{T}}(\mathbf{H}_{(k+1)} + \tfrac{N_0}{2}\,\mathbf{I})^{-1}\underline{H}_{(k)}\right). \tag{4}$$

We are now in a position to state the signal design problem: given received powers $\mathbf{P}$, and the QoS constraints that user k must achieve an SIR that is no less than β_k, find a correlation matrix $\mathbf{R}$ (or equivalently $\mathbf{H}$) with as small a rank (and hence bandwidth) as possible that will result in $\gamma_k \geq \beta_k$ for each $k = 1, 2, \cdots, K$ where γ_k is given in (4). We choose β_k to be some fraction $\eta_k \in (0, 1]$ of the maximum achievable SIR (that of a single-user channel), i.e., $\beta_k = 2\eta_k \frac{p_k}{N_0}$.

In [7] we obtain an algorithm for designing a correlation matrix that satisfies the QoS constraints and has small, if not minimal, rank. The algorithm is very simple to describe. Assume that the users are decoded in the increasing order of their indices (1 to K). The matrix $\mathbf{H}$ is constructed recursively as follows. Set $H_{KK} = p_K$. Given that $\mathbf{H}_{(k+1)}$ has been determined, $\mathbf{H}_{(k)}$ is obtained by specifying $\underline{H}_{(k)}$ to be

$$\underline{H}_{(k)} = \sqrt{\min\{(1-\eta_k)(\lambda_k + \tfrac{N_0}{2}), \lambda_k\}}\; p_k^{1/2}\underline{V}_k, \tag{5}$$

where λ_k is the smallest non-zero eigenvalue of $\mathbf{H}_{(k+1)}$ and $\underline{V}_k$ is its corresponding eigenvector.

There is, of course, the issue of choice of order of decoding. In a coded system erroneous feedback can be reduced to an acceptable level with good single-user codes so that, in principle, any ordering is fine [10]. Relative to some other orderings, it has been found that processing the users in decreasing order of $(1 - \eta_k)$ consistently does well at conserving bandwidth. The idea is to detect the users with more power to spare before those with little or no extra power.[1] Under a symmetric SIR constraint (i. e., $\gamma_k \geq \beta_{\text{sym}}$ for all k), this ordering rule is equivalent to decoding the users in descending order of p_k.

The algorithm is suboptimal in that it does not guarantee the minimal possible rank. However, it is shown in [7] that the above algorithm often yields a correlation matrix with a rank that is close to a (generally unachievable) lower bound on minimum rank (to be described below). In the best-case scenario, the minimum rank and our signal design strategy would yield the minimum processing gain of 1. In this case, it suffices for all users to employ the same signature waveform. In the worst-case scenario, the minimum rank required may be $M = K$ (e.g. when $\eta_k = 1$ for all k) in which case it is not possible to improve on the bandwidth required for Time or Frequency Division Multiple-Access (TDMA or FDMA).

---

[1] In an uncoded system, an ordering rule that directly mitigates error propagation effects, such as the decreasing order of the QoS requirements, is more appropriate.

The minimum rank is a function of the received powers and the SIR constraints but the problems of finding it and/or the corresponding correlation matrices (with minimum rank) that satisfy the QoS constraints are as yet unsolved. However, in [7] we find a (generally unachievable) lower bound by allowing a variable signaling rate $1/T$ (that depends on the received powers) and joint maximum-likelihood decoding at the receiver or rate-splitting multiple-access [11]. These strategies are however extremely difficult, if not impossible, to implement in practice.

Henceforth, the CWMA system that specifically uses the MSIR-DF receiver and the signals derived from the correlation matrix obtained according to the above recursive algorithm will be refered to as the BEMA system. The corresponding signal design will be referred to as the BEMA design.

III. SIGNAL DESIGNS FOR LINEAR RECEIVERS AND SUM-CAPACITY MAXIMIZATION

Recently, two other methods of signal design have been proposed in the literature.

In [4], Viswanath, Anantharam and Tse consider the linear MSIR receiver. For a given processing gain (and with or without a constraint on the average received powers), they obtain necessary and sufficient conditions under which there would exist powers and signature sequences that will allow the linear MSIR receiver to achieve a given set of target SIR constraints.

In [6], Viswanath and Anantharam consider the CWMA model with fixed users' received powers. For any fixed processing gain, they obtain signature sequences that maximize the sum capacity of the channel (this problem was solved earlier for the special case of equal fixed receive powers in [5]). As in BEMA, this approach does not employ power control, but unlike BEMA it focuses on a single design criterion (i. e., sum capacity) instead of a set of QoS criteria.

This section describes ways of interpreting the results of [4] and [6], optimal though they are in their respective contexts, within the context of a QoS-based approach to a combined design of signature waveforms and receiver to achieve bandwidth efficient multiple-access communications.

In addition to the notation used so far, we let the signature sequences (coefficients of the linear combination of basis function waveforms) of the users be arranged as unit-norm columns in an $M \times K$ matrix $\mathbf{A}$ (note that $\mathbf{R} = \mathbf{A}^\top\mathbf{A}$), and define $\mathbf{S} = \mathbf{I} + (\frac{N_0}{2})^{-1}\mathbf{P}^{1/2}\mathbf{A}^\top\mathbf{A}\mathbf{P}^{1/2}$.

A. Optimum Signal Design and Power Control for the Linear MSIR Receiver

Given a set of SIR constraints, the Viswanath-Anantharam-Tse (VAT) method of signal design determines whether or not there exists a set of signature sequences with processing gain M and a set of received powers (with or without an average-power constraint) such that each user meets its SIR constraint with a linear MSIR receiver. In this case, the SIR of the k^{th} user, given by $\gamma_k = 1/(\mathrm{S}^{-1})_{kk} - 1$, is required to be no less than β_k. Without loss of generality, assume that $\beta_1 \geq \beta_2 \geq \cdots \geq \beta_K$, and define the *effective bandwidth* of the k^{th} user as $e(\beta_k) \triangleq \beta_k(1+\beta_k)^{-1}$. There is a unique $k^* \in \{0, 1, \cdots M-1\}$ such that $(M-k^*)e(\beta_{k^*}) > \sum_{j=k^*+1}^{K} e(\beta_j) \geq (M-k^*)e(\beta_{k^*+1})$. A key result in [4] is as follows.

For a given processing gain $M < K$, there exists a $(\mathbf{P}, \mathbf{A})$ pair such that the users meet the SIR constraints and the average-power constraint, $K^{-1}\sum_{k=1}^{K} p_k \leq \overline{P}$, if and only if

$$M > \sum_{k=1}^{K} e(\beta_k) \quad \text{and} \quad \overline{P} \geq \frac{N_0/2}{K}\left(\sum_{i=1}^{k^*}\beta_i + \frac{(M-k^*)\sum_{i=k^*+1}^{K} e(\beta_i)}{M-k^*-\sum_{i=k^*+1}^{K}\beta_i}\right). \tag{6}$$

Without the power constraint, these two conditions reduce to only the first condition.

It is straightforward to view this design within the context of BEMA. In the BEMA design, we are given the users' received powers so that their average power is $\overline{P} = \frac{1}{K}\sum_{k=1}^{K} p_k$. Suppose that the BEMA signal design described Section II yields a correlation matrix with rank M_{BEMA}. Similarly, let M_{VAT} (subject to the average power constraint) and $M_{\text{VAT-U}}$ (unconstrained powers) denote the minimum required processing gains for the VAT signal design to meet the SIR constraints. In contrast to BEMA, the VAT and VAT-U designs have the additional freedom of choosing the most favorable power distribution, the former within the confines of an average power constraint and the latter without any such constraints. Consider now the following two examples.

Example 1: In this example the target SIR values of all users are assumed to be identical (i. e., we have a symmetric SIR constraint). Figure 2 plots the required processing gain versus SNR for fifteen users for several power distributions (see [12] for a justification of these distributions). The SNR is defined to be that of the weakest user, whose power is set at unity, so that $\text{SNR} = 10\log_{10}(1/N_0)$. The signaling interval is fixed at $T = 1.0$, and the symmetric SIR constraint is given by $\beta_{\text{sym}} = (\frac{N_0}{2})^{-1}$. Additionally, the figure also plots the FDMA upper bound and the lower bound mentioned in Section II.

It is clear from the figure that both BEMA and VAT/VAT-U lie within the upper and lower bounds, and that BEMA uniformly and significantly outperforms the VAT and VAT-U designs. The performance gains of BEMA are especially evident as the disparity in the users' powers is increased. Note also that there are situations where BEMA actually achieves the lower bound and is hence optimal in such cases.

It can be easily shown that for large enough SNR both VAT-U and BEMA are forced to require the same processing gain as FDMA (i. e., $M = K$). For VAT-U, this occurs whenever

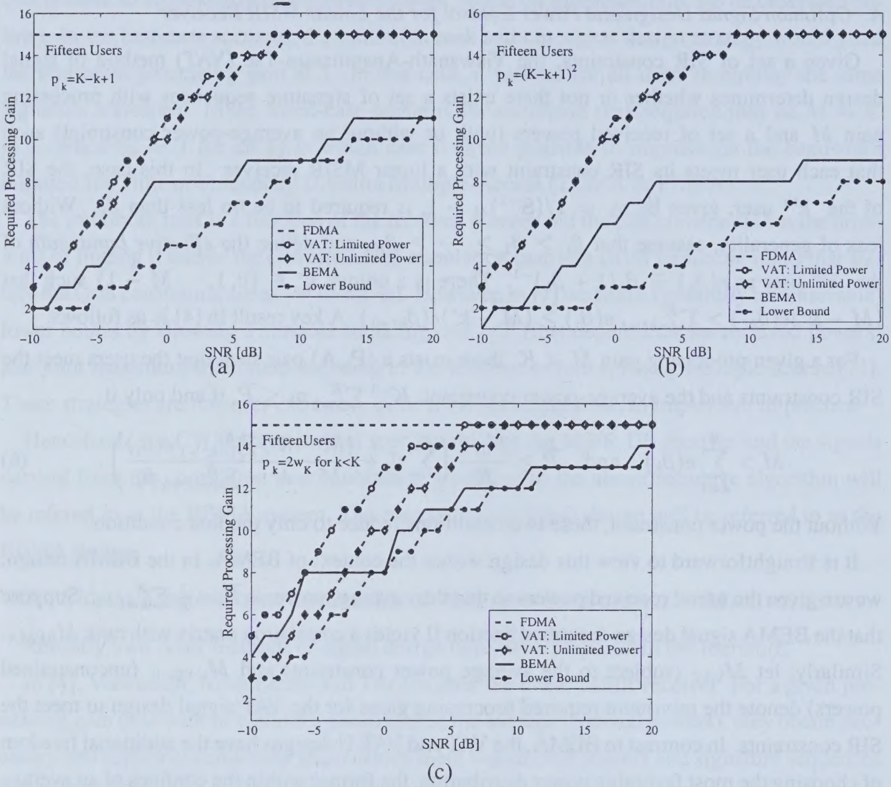


Fig. 2. Required processing gain versus SNR for fifteen users and several power disparities.

$\beta_{\text{sym}} > K - 1$. Thus, for the example at hand the processing gain of VAT-U is 15 whenever the SNR is greater than 8.451 dB. In contrast, we find experimentally that the processing gain of BEMA does not reach 15 in Figure 2 (a) until the SNR reaches 120.3 dB. Moreover, at this SNR, the lower bound on the processing gain is 14 which is itself close to that required by FDMA. Similarly, in Figure 2 (c), the processing gain of BEMA reaches 15 at an SNR of 41.2 dB, at which point the corresponding lower bound on the processing gain is again 14.

Example 2: In this example the SIR constraints are taken to be asymmetric. There are five users whose powers, in the case of BEMA, are fixed, and the signaling interval is once again set at $T = 1.0$. To establish a set of asymmetric SIR constraints, we randomly choose the SIR constraint of the k^{th} user, β_k, from the interval $[0, p_k/(N_0/2)]$ according to a uniform distribution. Given these constraints, the required processing gain of both BEMA and VAT/VAT-U are evaluated. In Figure 3 we plot a histogram of the processing gain of BEMA versus the processing gain of VAT/VAT-U over one thousand samples of the SIR constraints for users whose powers exhibit linear disparity; several values of SNR are considered. As was the case in the previ-

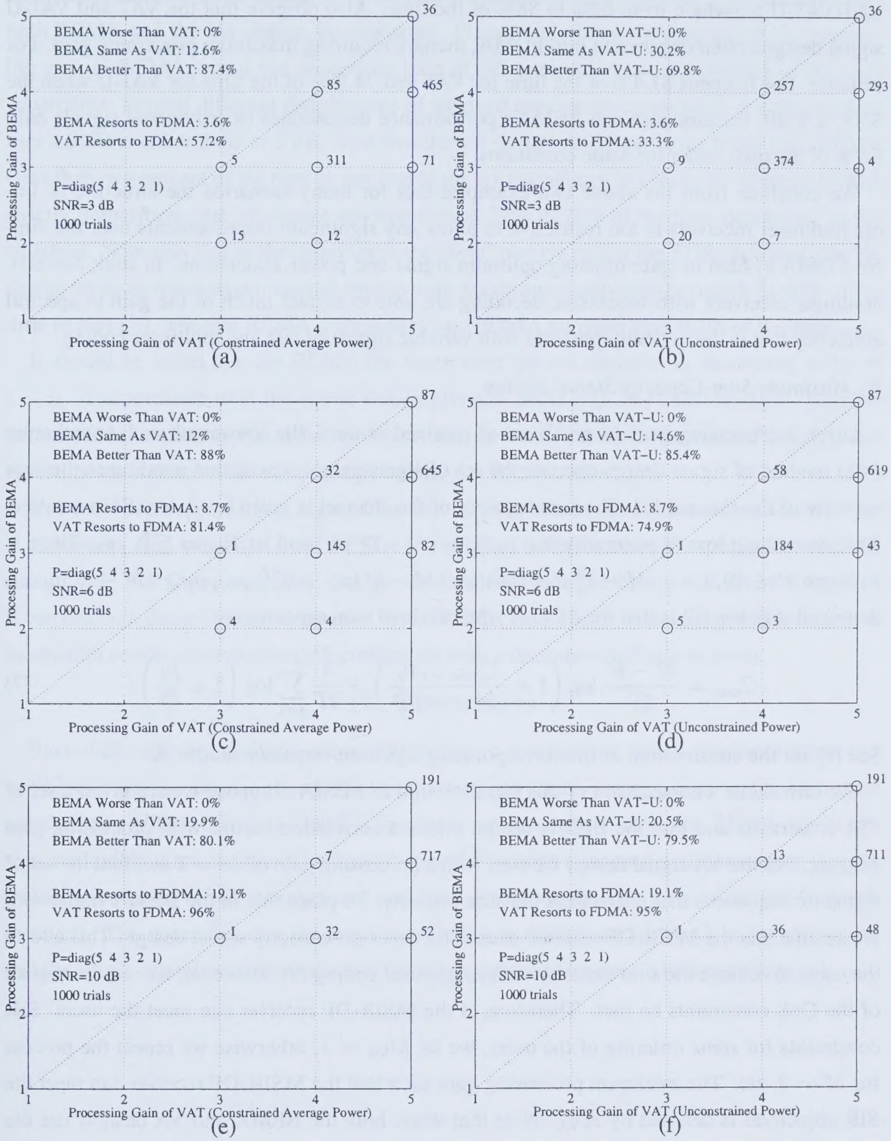


Fig. 3. Random non-symmetric QoS constraints with linearly disparate users.

ous example, BEMA once again dramatically outperforms the VAT and the VAT-U designs. Nowhere do VAT and VAT-U perform better than BEMA, while BEMA is strictly better than VAT/VAT-U anywhere from 69% to 88% of the time. Also observe that the VAT and VAT-U signal designs often degenerate into FDMA, thereby requiring maximum processing gain. For instance, this happens 81.4% of the time for VAT and 74.9 % of the time for VAT-U when the SNR is 6 dB. In stark contrast, BEMA's performance degenerates to orthogonal signals only 8.7% of the time under the same conditions.

We conclude from the above two examples that for many scenarios the structure of linear multiuser receivers is too restrictive to allow any significant improvements over the simple FDMA system in spite of using optimum signal and power allocations. In stark contrast, nonlinear receivers with successive decoding are able to extract much of the gain in spectral efficiency that is theoretically possible with variable signaling rate and multiuser coding.

B. Maximum Sum-Capacity Signal Design

Given a processing gain M and a set of received powers, the Viswanath and Anantharam (VA) method of signal design chooses the set of signature sequences that maximizes the sum capacity of the channel [6]. The sum capacity of the channel is given by $\frac{1}{2T}\log|\mathbf{S}|$ in nats/sec. Assume without loss of generality that $p_1 \geq p_2 \geq \cdots \geq p_K$, and let $P_{\text{tot}} = \sum_{k=1}^{K} p_k$. There is a unique $k^* \in \{0, 1, \cdots, M-1\}$ that satisfies $(M-k^*)p_{k^*} > \sum_{j=k^*+1}^{K} p_j \geq (M-k^*)p_{k^*+1}$. A central result in [6] is that for $M < K$, the maximal sum capacity is

$$C_{\text{sum}} = \frac{M-k^*}{2T}\log\left(1+\frac{\sum_{j=k^*+1}^{K} p_j}{(M-k^*)\frac{N_0}{2}}\right) + \frac{1}{2T}\sum_{j=1}^{k^*}\log\left(1+\frac{p_j}{\frac{N_0}{2}}\right). \qquad (7)$$

See [6] for the construction of the corresponding signature-sequence matrix $\mathbf{A}$.

We now make a comparison of this signal design to BEMA. Suppose we are given a set of SIR constraints and that the BEMA design yields a correlation matrix with processing gain M_{BEMA}. For the VA signal design we start with a processing gain of $M = 1$ and find the set of signature sequences that maximizes the sum capacity. To place this in the BEMA framework we assume that the MSIR-DF receiver is used for this sum-capacity signal design. This allows the users to achieve the sum capacity with single-user coding [9]. However, we require that all of the QoS constraints be met. Therefore, if the MSIR-DF receiver can meet the users' SIR constraints for *some* ordering of the users, we let $M_{\text{VA}} = 1$, otherwise we repeat the process for $M = 2$, etc. The minimum processing gain such that the MSIR-DF receiver can meet the SIR objectives is denoted by M_{VA}. Note that while both the BEMA and VA designs use the same type of receiver, the VA design has the extra freedom relative to BEMA of choosing the best permutation from all $K!$ orders in which users can be decoded.

Example 3: Consider a five-user channel where the received powers are fixed and the signaling interval is $T = 1.0$. To establish a set of asymmetric SIR constraints, we randomly choose the QoS of the k^{th} user to be the rate r_k [nats/sec] from the interval $[0, \frac{1}{2T}\log(1 + p_k/(N_0/2)]$ according to a uniform distribution.[2] From r_k it is easy to obtain the corresponding QoS constraint expressed as a SIR value, β_k. Given these constraints, the required processing gains of both BEMA and the VA design are evaluated. Figure 4 contains a histogram of the processing gain of BEMA versus the processing gain of VA over one thousand samples of the QoS constraints. Several different distributions of received powers are considered, and the weakest user's SNR is either 3 dB or 6 dB. Note that the VA design rarely performs better than BEMA (less than two percent of the time in any one of the six histograms in Figure 4), whereas BEMA strictly outperforms the VA design anywhere from 37% to 78% of the time depending on the situation. Note also that as the disparity in the users' powers and/or the SNR increases, the VA design tends to resort to orthogonal signals with maximum bandwidth (as much as 46% of the time in part (f)), whereas BEMA degenerates into FDMA no more than 0.3% of the time.

It should be noted that for BEMA the users were always decoded in decreasing order of $1 - \eta_k$ in accordance with the signal design given in Section II. By considering all possible permutations instead, as was done for the VA design in this example, the value of M_{BEMA} can sometimes be further reduced, although usually not significantly.

One must of course not lose sight of the fact that the VA design is optimal in the sense that it maximizes sum capacity. So if BEMA is evaluated in terms of sum capacity, it would be suboptimal. The above examples, though, do illustrate the point though that the sum-capacity is too coarse a design metric to be used in the BEMA framework where the problem is one of bandwidth conservation under QoS constraints with a decision-feedback receiver.

IV. CONCLUSIONS

Bandwidth-efficient multiple-access (BEMA) is a strategy for uplink communication that combines multiuser decision-feedback equalization and signal design under quality-of-service constraints. Within this QoS-based framework, we have shown that the BEMA signal design method requires much less bandwidth than the recently proposed optimum signal-designs methods for linear MSIR receivers and sum-capacity maximization (with a decision-feedback receiver). The constraint that the receiver be linear in the former approach is too restrictive, and the single-quantity maximization of the latter may be too coarse a design metric.

---

[2] Here we choose the QoS uniformly over the capacity of a single-user channel without any multiple-access interference, rather than uniformly over the valid range of SIR values. This measure of QoS is more in keeping with the original idea of BEMA as well as the sum-capacity maximization of the VA design. The conclusions are however similar for both cases.

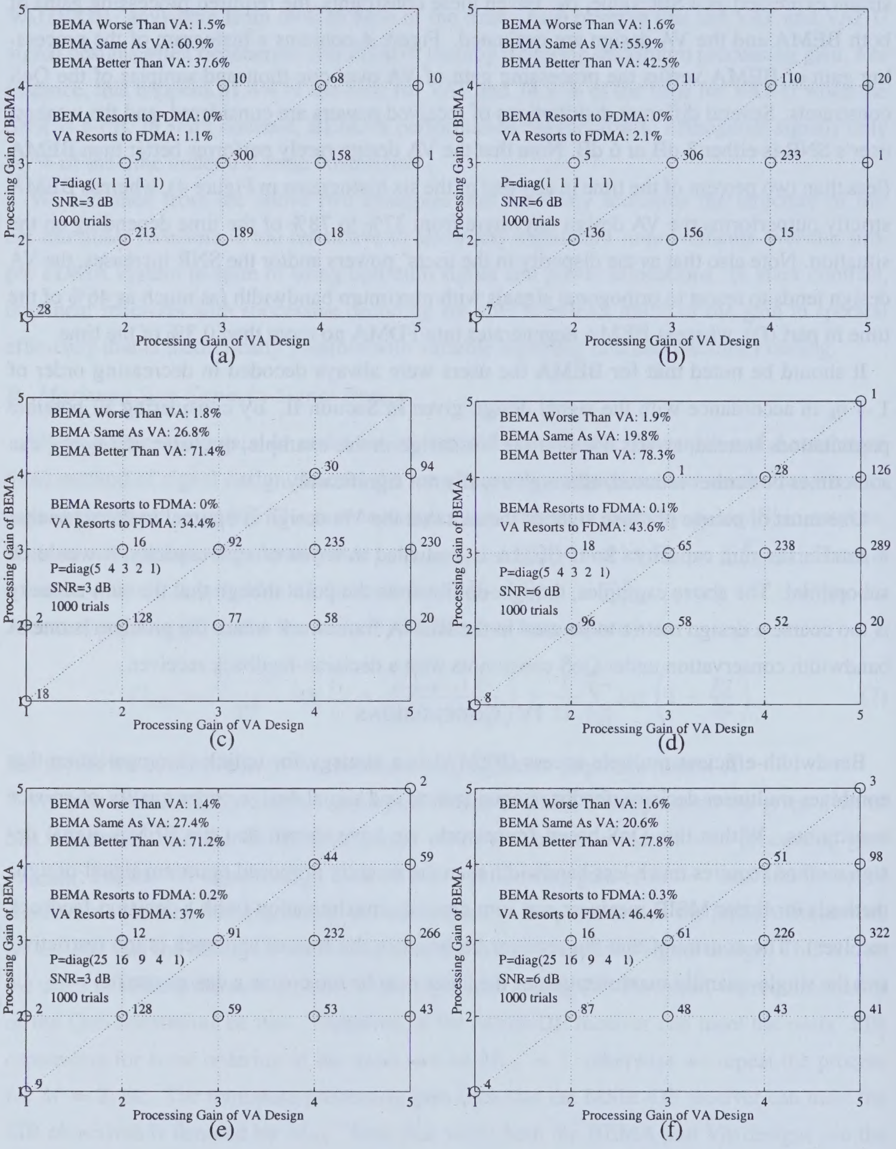


Fig. 4. Random non-symmetric QoS constraints for several user disparities and several SNR values.

REFERENCES

[1] T. Guess and M. K. Varanasi, " Onion Peeling for CDMA: Symmetric Capacity Under RMS-Bandwidth Constraints ," in *Proc. Thirtieth Annual Conference on Information Sciences and Systems*, Princeton, NJ, pp. 584–589, Mar. 1996.

[2] M. K. Varanasi and T. Guess, " Bandwidth-Efficient Multiple-Access via Signal Design for Decision-Feedback Receivers: Towards an Optimal Spreading-Coding Trade-Off ," in *Proc. IEEE Global Telecommun. Conf.: Commun. Theory Mini-Conf.*, Phoenix, AZ, pp. 159–165, Nov. 3-8 1997.

[3] M. K. Varanasi and T. Guess, " Bandwidth-Efficient Multiple-Access (BEMA): Signal Design with a Strict Bandwidth Criterion Under Quality-of-Service Constraints ," in *Proc. Intl. Symposium on Signals, Systems, and Electronics (ISSSE'98)* Pisa, Italy, Sep. 29-Oct. 2 1998.

[4] P. Viswanath, V. Anantharam, and D. Tse, "Optimal Sequences, Power Control and User Capacity of Synchronous CDMA Systems with Linear MMSE Multiuser Receivers," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1968–1983, Sept. 1999.

[5] M. Rupf and J. L. Massey, "Optimum Sequence Multisets for Synchronous Code-Division Multiple-Access Channels," *IEEE Trans. Inform. Theory*, vol. IT-40, pp. 1261–1266, Jul. 1994.

[6] P. Viswanath and V. Anantharam, "Optimal Sequences and Sum Capacity of Synchronous CDMA Systems," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1984–1991, Sept. 1999.

[7] M. K. Varanasi and T. Guess, " Bandwidth-Efficient Multiple Access (BEMA): A New Strategy Based on Signal Design for Multiuser Receivers Under Quality-of-Service Constraints ," Submitted *IEEE Trans. Commun.*, May 1998.

[8] T. Guess and M. K. Varanasi, " Multiuser Decision-Feedback Receivers for the General Gaussian Multiple-Access Channel ," in *Proc. Thirty-fourth Allerton Conf. Commun., Contr. and Comput.*, Allerton, IL, pp. 190–199, Oct. 1996.

[9] M. K. Varanasi and T. Guess, " Optimum Decision Feedback Multiuser Equalization with Successive Decoding Achieves the Total Capacity of the Gaussian Multiple-Access Channel ," in *Proc. Thirty-First Asilomar Conf. Signals, Systems, and Computers* , pp. 1405–1409, Nov. 1997.

[10] T. Guess and M. K. Varanasi, "Error Exponents for Maximum-Likelihood and Successive Decoders for the Gaussian CDMA Channel," *IEEE Trans. Inform. Theory*, vol. 46, pp. 1683–1691, Jul. 2000.

[11] B. Rimoldi and R. Urbanke, " A Rate-Splitting Approach to the Gaussian Multiple-Access Channel ," *IEEE Trans. Inform. Theory*, vol. IT-42, pp. 364–375, Mar. 1996.

[12] T. Guess and M. K. Varanasi, "Signal Design for Bandwidth-Efficient Multiple-Access Communications Based on Eigenvalue Optimization," *IEEE Trans. Inform. Theory*, vol. 46, pp. 2045–2058, Sept. 2000.

Combined CDMA Signal Design and Power Control for Decision-Feedback Receivers Subject to a Quality-of-Service Constraint

Tommy Guess
Department of Electrical Engineering, Univ. of Virginia, VA 22904-4743

38th Annual Allerton Conference on Communication, Control, and Computing, October 4-6, 2000

Abstract

We consider joint signal design and power control for the synchronous code-division multiple-access (CDMA) channel. All of the users are assigned the same quality-of-service (QoS) that is expressed in terms of a signal-to-interference ratio (SIR). A nonlinear decision-feedback receiver that successively detects the users is assumed. For a given processing gain, which is less than or equal to the number of users, we minimize the total received power over all signature sequences and power-control policies that meet the users' QoS constraint.

I. INTRODUCTION

Signal design for synchronous CDMA systems has received considerable attention over the past several years. As a function of processing gain, Rupf and Massey [1] designed signature sequences that maximize the sum capacity (i. e., the sum of the individual users' capacities) for the case when the users' received powers are all equal. A similar result was derived by Parsavand and Varanasi [2], but in the context of root mean-squared (RMS) bandwidth (as opposed to processing gain) and for the general case of unequal received powers. Viswanath and Anantharam [3] generalized the Rupf-Massey result from the case of equal received powers to the general case of unequal received powers. For a given set of received powers and QoS (i. e., SIR) constraints, bandwidth-efficient signal designs for decision-feedback receivers were derived by Varanasi and Guess [4] for strict bandwidth (or, equivalently, processing gain) and by Guess and Varanasi [5] for RMS bandwidth. As a function of processing gain, combined signature-sequence design and power control for linear receivers subject to SIR constraints was addressed by Viswanath, Anantharam and Tse [6], wherein optimal signature sequences and user power assignments were derived. The work at hand is most like this last reference, except that the linear receiver structure is replaced by a nonlinear one, and we focus on the case of a symmetric QoS constraint.

Given a K-user synchronous CDMA system with a QoS constraint assigned to each user and a processing gain N, our goal is to find the minimum total received power (i. e., the sum of the users' received powers) that is required for a nonlinear decision-feedback receiver to provide for each user his or her QoS constraint. It is of interest to note at the outset that for any set of

This work was supported in part by the Virginia Engineering Foundation.

signature sequences, then with enough power one can *always* guarantee the QoS constraint if a decision-feedback receiver is employed. This is in sharp contrast to linear receivers; as either the QoS constraint is increased or the processing gain is reduced, there will be inadmissible scenarios for which there does not exist a joint signal design and power distribution that allows a linear receiver to meet the QoS constraint for all users, even if the power is unlimited.

II. THE SYSTEM AND PROBLEM SETUP

Each of the K users employs pulse-amplitude modulation on an assigned unit-energy signature waveform, $u_i(t)$, to transmit the digital information symbol, x_i. A base station receives the superposition of the users' transmitted waveforms in symbol synchronism so that

$$y(t) = \sum_{i=1}^{K} x_i u_i(t) + n(t), \tag{1}$$

where $n(t)$ is additive white Gaussian noise with a two-sided, power spectral density of σ^2. The received signal-to-noise ratio (SNR) of the i-th user is given by $w_i = E[x_i^2]/\sigma^2$. Since it is this ratio that determines QoS performance, in what follows we can always assume that $\sigma^2 = 1$ so that received power is equal to received SNR. If we let $\{\phi_i(t)\}_{i=1}^{N}$ denote any orthonormal basis of the set of signature waveforms $\{u_i(t)\}_{i=1}^{K}$, where $N \le K$ is the processing gain, then it is possible to represent each signature waveform by a signature sequence of length N. Specifically, $u_i(t) = \sum_{j=1}^{N} s_{ij}\phi_j(t)$, with the real-valued signature sequence of user i given by the row vector $\underline{s}_i^{\mathsf{T}} = [s_{i1}\, s_{i2}\, \cdots\, s_{iN}]$ (the superscript T denotes the matrix transpose). Since the signature waveforms have unit energy, the signature sequences lie on the unit sphere. (We do not require that $s_{ij} = \pm\frac{1}{\sqrt{N}}$.)

Without loss of generality we assume the receiver front-end consists of a bank of filters matched to the users' signature waveforms, and that these outputs are sampled to yield the equivalent discrete-time channel model

$$\underline{Y} = \mathsf{S}^{\mathsf{T}}\mathsf{S}\underline{X} + \underline{V}, \tag{2}$$

where the $N \times K$ matrix S is given by $[\underline{s}_1\, \underline{s}_2\, \cdots\, \underline{s}_K]$ and $\underline{V}$ is a zero-mean Gaussian vector with covariance $\sigma^2\mathsf{S}^{\mathsf{T}}\mathsf{S}$. It is convenient to define the users' correlation matrix $\mathbf{R} = \mathsf{S}^{\mathsf{T}}\mathsf{S}$, a $K \times K$ positive semidefinite matrix with rank N. Note that if we are given any $K \times K$ correlation matrix with rank N, it is always possible to find a corresponding set of signature sequences with processing gain N. To see this, let $\mathbf{R} = \mathsf{U}^{\mathsf{T}}\mathbf{\Lambda}\mathsf{U}$ be a real-valued spectral decomposition. If the rank of $\mathbf{R}$ is N, then only N of its eigenvalues are greater than zero. This implies that we can express $\mathbf{R}$ as some $\mathsf{S}^{\mathsf{T}}\mathsf{S}$ where S is an $N \times K$ matrix whose columns are the users' signature sequences.

A. The Receiver

First consider a linear receiver parameterized by the $K \times K$ matrix $\mathbf{F}$. The i-th row of $\mathbf{F}$, which we denote by $\underline{f}_i^{\mathsf{T}}$, is used by the receiver to evaluate $z_i = \underline{f}_i^{\mathsf{T}} \underline{Y}$. This scalar statistic is used to detect symbol x_i. Special cases include the conventional receiver ($\mathbf{F} = \mathbf{I}$, the identity matrix) and the decorrelator ($\mathbf{F} = \mathbf{R}^{-1}$, whenever $\mathbf{R}$ is full rank) [7]. We are most interested in the linear receiver that has the property of maximizing the SIRs of the users. Such a receiver is given by $\mathbf{F} = (\mathbf{W}^{1/2}\mathbf{R}\mathbf{W}^{1/2} + \mathbf{I})^{-1}\mathbf{W}^{1/2}$, where we have defined $\mathbf{W} = \mathrm{diag}(w_1, w_2, \ldots, w_K)$. Note that if $\mathbf{F}$ is pre-multiplied by any diagonal matrix with non-zero diagonal entries, it retains its maximum-SIR property, and for the appropriate scaling matrix it will also be the minimum mean-squared error (MMSE) linear receiver [8]. The resulting SIR for user i is given by

$$\mathsf{SIR}_i = \frac{1}{((\mathbf{W}^{1/2}\mathbf{R}\mathbf{W}^{1/2} + \mathbf{I})^{-1})_{ii}} - 1. \tag{3}$$

This quantity represents the ratio of the effective received power of the i-th user after linear processing, $w_i((\mathbf{FR})_{ii})^2$, to the sum of the filtered noise power, $\sigma^2 \underline{f}_i^{\mathsf{T}} \mathbf{R} \underline{f}_i$, and the multiple-access interference contributed by the other users, $\sum_{j \neq i} w_j ((\mathbf{FR})_{ij})^2$.

The linear receiver structure can be generalized to the nonlinear decision-feedback receiver depicted in Figure 1 [9]. It is parameterized by $(\mathbf{F}, \mathbf{B})$, where both $\mathbf{F}$ and $\mathbf{B}$ are $K \times K$ matrices with $\mathbf{B}$ constrained to be strictly lower triangular. The users are detected successively. To detect the i-th user, the receiver calculates $z_i = \underline{f}_i^{\mathsf{T}} \underline{Y} - \sum_{j=1}^{i-1} B_{ij} \hat{x}_j$ where $\{\hat{x}_j\}_{j=1}^{i-1}$ indicates the symbol decisions for the users that have already been detected. We will assume that the symbol decisions are always correct so that the feedback is perfect. Such an assumption can be justified in a coded system [10], but even in uncoded systems the effects of error propagation can for the most part be mitigated when the users are decoded in decreasing order of SIR [11]. No matter the choice of $\mathbf{F}$, under the perfect-feedback assumption the best one can do as far as maximizing SIR is to let $\mathbf{B}$ equal the strictly lower-triangular part of $\mathbf{FR}$. Moreover, given the order in which the users are to be processed, there exists a decision-feedback receiver that simultaneously maximizes the SIR of each user [12]. For the case when the users are detected in the order indicated by their indices (we can always permute the users to get other orderings),

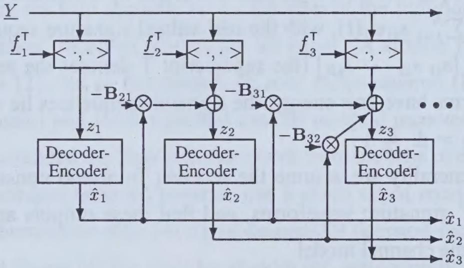


Fig. 1. A multiple-access channel decision-feedback receiver structure.

we let $\mathbf{W}^{1/2}\mathbf{R}\mathbf{W}^{1/2} + \mathbf{I} = \mathbf{L}^{\mathsf{T}}\mathbf{D}\mathbf{L}$ be the unique Cholesky decomposition where $\mathbf{L}$ is lower triangular with unit diagonal elements and $\mathbf{D}$ is a diagonal matrix. Then $\mathbf{F} = (\mathbf{L}^{\mathsf{T}})^{-1}\mathbf{W}^{1/2}$. The users' SIR values are easily determined by the diagonal entries of $\mathbf{D}$ since $d_i - 1$ is the SIR of the i-th user, where we have let $\mathbf{D} = \text{diag}(d_1, d_2, \ldots, d_K)$. For ease of discussion we refer to the diagonal elements of $\mathbf{D}$ as Cholesky values.

B. Problem Statement

We are now in a position to state the problem considered in this paper. For a given processing gain less than or equal to the number of users, minimize the total received power over all sets of signature sequences and user power assignments such that a decision-feedback receiver provides for all users the SIR constraint. Thus we seek the solution to

$$\min_{\mathbf{W},\mathbf{R}} \{\text{trace}(\mathbf{W}) : \text{rank}(\mathbf{R}) = N, \{d_i \geq \mathsf{SIR} + 1\}_{i=1}^{K}\}, \tag{4}$$

where $\{d_i\}_{i=1}^{K}$ are the Cholesky values of

$$\mathbf{A} = \mathbf{I} + \mathbf{W}^{1/2}\mathbf{R}\mathbf{W}^{1/2}. \tag{5}$$

It is implicitly understood that $\mathbf{W}$ is a diagonal positive-definite matrix and $\mathbf{R}$ is positive semidefinite with unity-valued diagonal elements. Note that the rank of $\mathbf{R}$ is equal to the number of $\mathbf{A}$'s eigenvalues that are different from one. Also note in (4) that $\text{trace}(\mathbf{W})$ can be replaced by $\sum_{i=1}^{K} \lambda_i - K$. Since $K - N$ of these eigenvalues are unity, we are essentially trying to minimize the sum of the remaining N eigenvalues. It is this step that complicates matters, because it turns out that not all choices for the remaining eigenvalues are valid. To see that this is the case, note that the determinant of $\mathbf{A}$ is equal to product of its eigenvalues as well as the product of its Cholesky values. This imposes a constraint on our eigenvalue choices. In the following section, we determine the set of constraints on the eigenvalues of we which we must be mindful as we attempt to minimize their sum.

III. THE RELATIONSHIP BETWEEN EIGENVALUES AND CHOLESKY VALUES

In order to solve the joint signal-design and power-control problem it is necessary to find a connection between the Cholesky values and eigenvalues of positive-definite matrices. Our concern with the Cholesky values stems from their explicit connection to the users' SIR constraint, while the eigenvalues are important since they determine both the required processing gain and the total received power.

A. Majorization

To identify the constraints on the eigenvalues and Cholesky values of $\mathbf{A}$ we must work with the concept of majorization. Given two length-n sequences, majorization is defined as follows.[1]

Definition 1: Let $a_1 \geq a_2 \geq \cdots \geq a_n$ and $b_1 \geq b_2 \geq \cdots \geq b_n$ be two ordered sequences of length n. We say that $\underline{a} = (a_1, a_2, \ldots, a_n)$ majorizes $\underline{b} = (b_1, b_2, \ldots, b_n)$ if $\sum_{i=1}^{k} a_i \geq \sum_{i=1}^{k} b_i$ for all $k = 1, 2, \ldots, n$, and with equality when $k = n$. In notation we write $\underline{a} > \underline{b}$.

It is simple to show that the condition in Def. 1 may be replaced by $\sum_{i=1}^{k} a_{n+1-i} \leq \sum_{i=1}^{k} b_{n+1-i}$ for all $k = 1, 2, \ldots, n$, and with equality when $k = n$. It is sometimes easier to work with this equivalent representation.

B. Majorization and the Cholesky Decomposition

Consider now an arbitrary $n \times n$ positive-definite matrix $\mathbf{A}$. It is well known that there exists a spectral decomposition of such a matrix given by $\mathbf{A} = \mathbf{U}^{\dagger}\mathbf{\Lambda}\mathbf{U}$ where $\mathbf{U}$ is a unitary matrix and $\mathbf{\Lambda}$ is a diagonal matrix whose diagonal entries are the eigenvalues of $\mathbf{A}$ (the superscript $\dagger$ denotes the Hermitian transpose). But $\mathbf{A}$ also has another common diagonalization known as the Cholesky factorization. It is given by $\mathbf{A} = \mathbf{L}^{\dagger}\mathbf{D}\mathbf{L}$, where $\mathbf{L}$ is a lower-triangular matrix with unity-valued diagonal elements and $\mathbf{D}$ is a diagonal matrix [13]. We represent the Cholesky values by $\underline{d} = (d_1, d_2, \ldots, d_n)$ and the eigenvalues by $\underline{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_n)$. Furthermore, without any loss in generality we can assume that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. In the following lemma we are concerned with the logarithms of these values and for notational felicity we let $\log(\underline{d})$ denote $(\log(d_1), \log(d_2), \ldots, \log(d_n))$ and $\log(\underline{\lambda})$ denote $(\log(\lambda_1), \log(\lambda_2), \ldots, \log(\lambda_n))$.

Lemma 1: The eigenvalues, $\underline{\lambda}$, and Cholesky values, $\underline{d}$, of a positive-definite matrix satisfy $\log(\underline{\lambda}) > \log(\underline{d})$.

Proof: The statement is obviously true for $n = 1$. We now suppose that it is true for $n - 1$ and then show that this implies it is also true for n. Partition the Cholesky factorization of $\mathbf{A}$ as

$$\mathbf{A} = \begin{bmatrix} 1 & (\mathbf{L}_{2:n,1})^{\dagger} \\ \underline{0} & \mathbf{L}_{2:n,2:n}^{\dagger} \end{bmatrix} \begin{bmatrix} d_1 & \underline{0}^{\mathrm{T}} \\ \underline{0} & \mathbf{D}_{2:n,2:n} \end{bmatrix} \begin{bmatrix} 1 & \underline{0}^{\mathrm{T}} \\ \mathbf{L}_{2:n,1} & \mathbf{L}_{2:n,2:n} \end{bmatrix}, \tag{6}$$

where for a matrix $\mathbf{B}$ the notation $\mathbf{B}_{a:b,c:d}$ denotes the submatrix formed by retaining only those rows in $\{a, a+1, \ldots, b\}$ and those columns in $\{c, c+1, \ldots, d\}$. It is easily observed that the submatrix $\mathbf{A}_{2:n,2:n}$ is equal to $\mathbf{L}_{2:n,2:n}^{\dagger}\mathbf{D}_{2:n,2:n}\mathbf{L}_{2:n,2:n}$. This, incidentally, is its Cholesky factorization. We identify the submatrix $\mathbf{A}_{2:n,2:n}$ by $\hat{\mathbf{A}}$. The Cholesky values of $\hat{\mathbf{A}}$ are $\underline{\hat{d}} = (d_2, d_3, \ldots, d_n)$, and we let $\hat{d}_{j_1} \geq \hat{d}_{j_2} \geq \cdots \geq \hat{d}_{j_{n-1}}$ be an ordering of these values. $\hat{\mathbf{A}}$ also has an associated spectral decomposition given by $\hat{\mathbf{A}} = \hat{\mathbf{U}}^{\dagger}\hat{\mathbf{\Lambda}}\hat{\mathbf{U}}$ where, without loss of generality it is assumed that $\hat{\lambda}_1 \ge \hat{\lambda}_2 \ge \cdots \ge \hat{\lambda}_{n-1}$. Now

---

[1] The reader should be aware that another definition of majorization reverses the direction of the inequality symbol (e. g., [13]).

$$\begin{aligned}\sum_{i=1}^{n} \log(\lambda_i) &= \sum_{i=1}^{n} \log(d_i) &(7)\\ &= \log(d_1) + \sum_{i=1}^{n-1} \log(\hat{d}_{j_i}) &(8)\\ &= \log(d_1) + \sum_{i=1}^{n-1} \log(\hat{\lambda}_i), &(9)\end{aligned}$$

where the first and last equations are a consequence of the determinant of a positive-definite matrix being equal to both the product of its eigenvalues and the product of its Cholesky values. Thus we have that

$$\log(d_1) = \sum_{i=1}^{n} \log(\lambda_i) - \sum_{i=1}^{n-1} \log(\hat{\lambda}_i). \tag{10}$$

Now because $\mathbf{A}$ is a bordered matrix relative to the submatrix $\hat{\mathbf{A}}$, we know that their respective eigenvalues obey the interlacing property given by $\lambda_1 \ge \hat{\lambda}_1 \ge \lambda_2 \ge \hat{\lambda}_2 \ge \cdots \ge \hat{\lambda}_{n-1} \ge \lambda_n$ [13]. We conclude therefore that $\log(d_1)$ falls somewhere within the closed interval $[\log(\lambda_n), \log(\lambda_1)]$. There are n possible cases as to where d_1 falls in relation to those values in $\hat{\underline{d}}$, Take case l), where $\hat{d}_{j_{l-1}} \ge d_1 \ge \hat{d}_{j_l}$. We already know by interlacing and the induction hypothesis that $\sum_{i=1}^{s} \log(\lambda_i) \ge \sum_{i=1}^{s} \log(\hat{\lambda}_i) \ge \sum_{i=1}^{s} \log(\hat{d}_{j_i})$ for $s = 1, 2, \ldots, l$. So we ask whether or not it is possible for $\sum_{i=1}^{l-1} \log(\hat{d}_{j_i}) + \log(d_1) > \sum_{i=1}^{l} \log(\lambda_i)$, which would imply that majorization does not hold. This occurs if and only if $\sum_{i=l+1}^{n} \log(\lambda_i) > \sum_{i=l}^{n-1} \log(\hat{d}_{j_i})$. But the latter cannot be true without leading to a contradiction, since it also implies that $\sum_{i=l}^{n-1} \log(\hat{\lambda}_i) > \sum_{i=l}^{n-1} \log(\hat{d}_{j_i})$ which is cannot be true if $\log(\hat{\underline{\lambda}}) > \log(\hat{\underline{d}})$. Similarly, it can be shown for each $s = l+1, l+2, \ldots, n$ that $\sum_{i=1}^{s} \log(\lambda_i) \ge \log(d_1) + \sum_{i=1}^{s-1} \log(\hat{d}_{j_i})$. Thus we must conclude that $\log(\underline{\lambda}) > \log(\underline{d})$ since the elements of $\hat{\underline{d}}$ are some permutation of $(d_2, d_3, \ldots, d_n)$. ■

Now that we have a very specific statement that relates the eigenvalues and Cholesky values of positive-definite matrices, it is of interest to consider the converse relationship. That is, given two length-n sequences, one of which majorizes the other, we ask whether or not there exists a corresponding positive-definite matrix whose eigenvalues and Cholesky values are determined by the sequences. This is in fact the case, as we shall demonstrate, but first we must give a majorization lemma.

Lemma 2: Suppose that $a_1 \ge a_2 \ge \cdots \ge a_n$ and $b_1 \ge b_2 \ge \cdots \ge b_n$ with $\underline{a} > \underline{b}$. Then for each $k \in \{1, 2, \ldots, n\}$ there exists an $n-1$ length sequence $(c_1, c_2, \ldots, c_{k-1}, c_{k+1}, \ldots, c_n)$ that satisfies an interlacing constraint and a majorization constraint:

(i) $a_1 \ge c_1 \ge a_2 \ge c_2 \ge \cdots \ge a_{k-1} \ge c_{k-1} \ge a_k \ge c_{k+1} \ge a_{k+1} \ge \cdots \ge c_n \ge a_n$

(ii) $(c_1, c_2, \ldots, c_{k-1}, c_{k+1}, \ldots, c_n) > (b_1, b_2, \ldots, b_{k-1}, b_{k+1}, \ldots, b_n)$.

It may be recognized that this lemma is a generalization of Lemma 4.3.28 in [13] in that any element of $\underline{b}$ may be removed, not just the last or first element. Moreover, in contrast to the proof in [13] our proof of this lemma, though not included here due to space constraints, provides an explicit algorithm for constructing the desired sequence.

The following lemma can now be proved without much difficulty.

Lemma 3: Given n-length positive-valued sequences $\underline{\lambda}$ and $\underline{d}$ such that $\log(\underline{\lambda}) \succ \log(\underline{d})$, let $d_{j_1}, d_{j_2}, \ldots, d_{j_n}$ denote a desired permutation of the elements in $\underline{d}$. Then there exists a real-valued positive-definite matrix $\mathbf{A}$ whose eigenvalues and Cholesky values are those elements of $\underline{\lambda}$ and $\underline{d}$, respectively. Moreover, if the Cholesky factorization of $\mathbf{A}$ is given by $\mathbf{A} = \mathbf{L}^{\mathsf{T}}\mathbf{D}\mathbf{L}$, then we have that $\mathbf{D} = \operatorname{diag}(d_{j_1}, d_{j_2}, \ldots, d_{j_n})$.

Proof: Assume that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ and $d_1 \geq d_2 \geq \cdots \geq d_n$ without any loss in generality. Suppose the lemma statement is true for sequences of length $n-1$, then we shall show that this implies its truth for sequences of length n. Since the lemma is obviously true for $n = 1$, we will have proved by induction its veracity for all n.

Let $k = j_1$, the first index of our desired ordering of the Cholesky values. Then by Lemma 2 we know that there exists a length $n-1$ sequence $(\gamma_1, \gamma_2, \ldots, \gamma_{k-1}, \gamma_k, \ldots, \gamma_n)$ such that

$$(i) \quad \log(\lambda_1) \geq \gamma_1 \geq \log(\lambda_2) \geq \gamma_2 \geq \cdots \geq \log(\lambda_{k-1}) \geq \gamma_{k-1} \geq \log(\lambda_k) \geq \gamma_{k+1} \geq \cdots \geq \gamma_n \geq \log(\lambda_n) \tag{11}$$

$$(ii) \quad (\gamma_1, \ldots, \gamma_{k-1}, \gamma_{k+1}, \ldots, \gamma_n) \succ (\log(d_1), \ldots, \log(d_{k-1}), \log(d_{k+1}), \ldots, \log(d_n)). \tag{12}$$

The induction hypothesis implies there exists an $(n-1) \times (n-1)$ matrix $\mathbf{B}$ with spectral decomposition $\mathbf{B} = \hat{\mathbf{U}}^{\mathsf{T}}\hat{\mathbf{\Lambda}}\hat{\mathbf{U}}$ and Cholesky decomposition $\mathbf{B} = \hat{\mathbf{L}}^{\mathsf{T}}\hat{\mathbf{D}}\hat{\mathbf{L}}$, where $\hat{\mathbf{\Lambda}} = \operatorname{diag}(e^{\gamma_1}, e^{\gamma_2}, \ldots, e^{\gamma_{k-1}}, e^{\gamma_{k+1}}, \ldots, e^{\gamma_n})$, $\hat{\mathbf{D}} = \operatorname{diag}(d_1, d_2, \ldots, d_{k-1}, d_{k+1}, \ldots, d_n)$. Use this $\mathbf{B}$ to create an $n \times n$ bordered matrix

$$\hat{\mathbf{A}} = \begin{bmatrix} \alpha & \underline{y}^{\mathsf{T}} \\ \underline{y} & \hat{\mathbf{\Lambda}} \end{bmatrix}. \tag{13}$$

By the converse to the interlacing-eigenvalue theorem, α and $\underline{y}$ can always be chosen such that the eigenvalues of $\hat{\mathbf{A}}$ are $\lambda_1, \lambda_2, \ldots, \lambda_n$ [13]. Let α and $\underline{y}$ be chosen as such, then create

$$\mathbf{A} = \begin{bmatrix} 1 & \underline{0}^{\mathsf{T}} \\ \underline{0} & \hat{\mathbf{U}}^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \alpha & \underline{y}^{\mathsf{T}} \\ \underline{y} & \hat{\mathbf{\Lambda}} \end{bmatrix} \begin{bmatrix} 1 & \underline{0}^{\mathsf{T}} \\ \underline{0} & \hat{\mathbf{U}} \end{bmatrix}. \tag{14}$$

To see that $\mathbf{A}$ satisfies the lemma requirements, first note that it is a unitary transformation of $\hat{\mathbf{A}}$, so it has the same eigenvalues as $\hat{\mathbf{A}}$. Secondly, observe that $\mathbf{A}$ has the Cholesky factorization

$$\mathbf{A} = \begin{bmatrix} 1 & \underline{y}^{\mathsf{T}}\hat{\mathbf{U}}\hat{\mathbf{L}}^{-1}\hat{\mathbf{D}}^{-1} \\ \underline{0} & \hat{\mathbf{L}}^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \alpha - \underline{y}^{\mathsf{T}}\hat{\mathbf{\Lambda}}^{-1}\underline{y} & \underline{0}^{\mathsf{T}} \\ \underline{0} & \hat{\mathbf{D}} \end{bmatrix} \begin{bmatrix} 1 & \underline{0}^{\mathsf{T}} \\ \hat{\mathbf{D}}^{-1}(\hat{\mathbf{L}}^{\mathsf{T}})^{-1}\hat{\mathbf{U}}^{\mathsf{T}}\underline{y} & \hat{\mathbf{L}} \end{bmatrix}. \tag{15}$$

Thus the Cholesky values of $\mathbf{A}$ are $d_1, d_2, \ldots, d_{k-1}, d_{k+1}, \ldots, d_n, \alpha - \underline{y}^{\mathsf{T}} \hat{\Lambda}^{-1} \underline{y}$. But since $\sum_{i=1}^{n} \log(d_i) = \sum_{i=1}^{n} \log(\lambda_i)$, we see that $d_k = \alpha - \underline{y}^{\mathsf{T}} \hat{\Lambda}^{-1} \underline{y}$ must be true. Therefore the constructed $\mathbf{A}$ has the desired eigenvalues and Cholesky values. ■

Together Lemmas 1 and 3 provide the inroads necessary for solving the joint signal-design and power-control problem proposed in Section II, for they show that majorization is the precise relationship between the eigenvalues and Cholesky values of a positive-definite matrix. It is of interest to recall that majorization also provides a precise relationship for Hermitian-symmetric matrices. The eigenvalues always majorize the diagonal elements of matrices from this class. Conversely, given that $\underline{a} > \underline{b}$, then there exists a Hermitian-symmetric matrix with eigenvalues $\underline{a}$ and diagonal elements $\underline{b}$ [13]. It is matrix diagonalization that provides the catalyst for both this result and Lemmas 1 and 3. The former relies on the spectral decomposition and is seen to give more details concerning the equality of the trace of a matrix and the sum of its eigenvalues. The latter relies on the Cholesky decomposition and is seen to provide more details concerning the equality of the product of the eigenvalues of a matrix and the product of its Cholesky values. It is in the former sense that majorization has played a key role other signal designs such as the maximization of sum capacity [2] [3] and joint signal-design and power-control for linear receivers subject to SIR constraints [6].

IV. THE MAIN RESULT

We are now in a position to solve the joint signal-design and power-control problem for decision-feedback receivers with a QoS constraint. To see this, let us suppose for a processing gain N that we have designed some policy $(\mathbf{S}, \mathbf{W})$ such that $\mathbf{S}$ is $N \times K$ and all of the QoS constraints are satisfied. Recalling $\mathbf{A}$ as defined in (5), we know that it is a positive-definite matrix with N eigenvalues greater than one and $K - N$ eigenvalues equal to one. We choose to label these in descending order as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N > 1$ where $\lambda_i = 1$ for $i = N+1$ through K. The total received SNR is thus $W_{\text{req}} = \sum_{i=1}^{N} \lambda_i - N$. The Cholesky values of $\mathbf{A}$ meet the QoS constraint $d_i \geq d \triangleq \mathsf{SIR} + 1$ for each i. The inequality can be replaced by equality since "beating" the SIR thresholds can never decrease the required power. Lemmas 1 and 3 allow us to restate the problem as

$$W_{\min} = \min \left\{ \sum_{i=1}^{N} \lambda_i - N : (\log(\lambda_1), \ldots, \log(\lambda_N), 0, \ldots, 0) > (\log(d), \ldots, \log(d)) \right\}. \quad (16)$$

Solving this equivalent problem provides the following proposition.

Proposition 1: Consider a K-user synchronous CDMA channel with a pre-specified QoS $d = \mathsf{SIR} + 1$. If an SIR-maximizing decision-feedback receiver is employed for some arbitrary ordering of the users, then the minimum required total received SNR over all QoS-achieving signal-design/power-control policies is given by

$$W_{\min} = N(\mathsf{SIR}+1)^{K/N} - N. \tag{17}$$

Proof: By the inequality relating the arithmetic and geometric means we know that $\frac{1}{N}\sum_{i=1}^{N}\lambda_i \geq (\prod_{i=1}^{N}\lambda_i)^{1/N}$ with equality only when all of the terms are equal. Now $\prod_{i=1}^{N}\lambda_i = \prod_{i=1}^{K} d_i$, and this implies that the received SNR is minimized by letting $\lambda_1 = \lambda_2 = \cdots = \lambda_N = d^{K/N}$ since these eigenvalues also satisfy the majorization constraint. Hence we find that the minimum required total SNR is $Nd^{K/N} - N$. ■

Since the logarithms of these majorize the logarithms of the Cholesky values, we know by Lemma 3 that there exists a real-valued, positive-definite matrix that possesses these eigenvalues and Cholesky values. Call this matrix $\mathbf{A}$. Then we have that the power-control policy is $\mathbf{W} = \mathrm{diag}(\mathbf{A} - \mathbf{I})$, and the correlation matrix of the signal design is $\mathbf{R} = \mathbf{W}^{-1/2}(\mathbf{A} - \mathbf{I})\mathbf{W}^{-1/2}$. Since only $N - K$ of $\mathbf{A}$'s eigenvalues are greater than one, we know that the rank of $\mathbf{R}$ is N. It can therefore be written as a symmetric product $\mathbf{R} = \mathbf{S}^{\mathsf{T}}\mathbf{S}$ where $\mathbf{S}$ is the $N \times K$ signal-design matrix.

V. AN EXAMPLE

We now illustrate the case where there are twenty users with a symmetric SIR constraint. As the processing gain is varied from one to twenty, we calculate the required average received SNR (that is, the total received SNR divided by the number of users) for optimal signal-design/power-control for both the linear (cf. [6]) and nonlinear (cf. Proposition 1) receivers. For the linear receiver the total received SNR must satisfy $W_{\mathrm{lin}} > 1 + \frac{1}{\mathsf{SIR}} - \frac{K}{N}$ [6]. In the cases pictured in Figure 2, it is observed that most processing gains produce inadmissible scenarios for the linear receiver. That is, no signal-design/power-allocation exists such that a linear receiver can achieve the QoS for every user, even if the power is unbounded. For unbounded power a valid allocation is possible only when $\mathsf{SIR} \leq \frac{N}{K-N}$. In contrast, the decision-feedback receiver can meet the QoS constraints for all processing gains, though for small ranks it can become exceedingly large.

Consider specifically the $\mathsf{SIR} = 10$ dB graph. For the full-rank system where the processing gain is equal to the number of users, the required average SNR is the same for both the linear and nonlinear receivers, namely 10 dB. This is achieved for both receivers with orthogonal signature sequences (e. g., FDMA). If the average received power increased to 13.25 dB, the processing gain required by the linear system can be reduced from 20 to 19, but any further reduction in processing gain is not possible without violating the QoS constraints, even if the power is unlimited. For the decision-feedback receiver, however, an average received power of 13.25 dB allows us to reduce the processing gain all the way down to 14. And with 20.0 dB we need only a processing gain of 9.

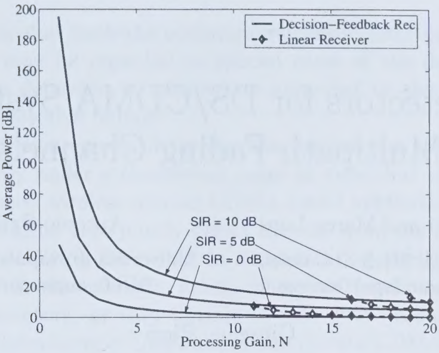


Fig. 2. Required average power versus processing gain for twenty users.

VI. CONCLUSION

When the users in a synchronous CDMA system are guaranteed a QoS constraint, optimal joint signal-design and power-control for nonlinear decision-feedback receivers provides substantial savings in processing gain compared to optimal joint signal-design and power-control for linear receivers.

REFERENCES

[1] M. Rupf and J. L. Massey, "Optimum Sequence Multisets for Synchronous Code-Division Multiple-Access Channels," *IEEE Trans. Inform. Theory*, vol. IT-40, pp. 1261–1266, Jul. 1994.

[2] D. Parsavand and M. K. Varanasi, " RMS Bandwidth Constrained Signature Waveforms that Maximize the Total Capacity of PAM-Synchronous CDMA Channels ," *IEEE Trans. Commun. Theory*, vol. COM-44, pp. 65–75, Jan. 1996.

[3] P. Viswanath and V. Anantharam, "Optimal Sequences and Sum Capacity of Synchronous CDMA Systems," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1984–1991, Sept. 1999.

[4] M. K. Varanasi and T. Guess, " Bandwidth-Efficient Multiple-Access (BEMA): Signal Design with a Strict Bandwidth Criterion Under Quality-of-Service Constraints ," in *Proc. Intl. Symposium on Signals, Systems, and Electronics (ISSSE'98)* Pisa, Italy, Sep. 29-Oct. 2 1998.

[5] T. Guess and M. K. Varanasi, "Signal Design for Bandwidth-Efficient Multiple-Access Communications Based on Eigenvalue Optimization," *IEEE Trans. Inform. Theory*, vol. 46, pp. 2045–2058, Sept. 2000.

[6] P. Viswanath, V. Anantharam, and D. Tse, "Optimal Sequences, Power Control and User Capacity of Synchronous CDMA Systems with Linear MMSE Multiuser Receivers," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1968–1983, Sept. 1999.

[7] R. Lupas and S. Verdu, "Linear Multiuser Detectors for Synchronous Code-Division Multiple-Access Channels," *IEEE Trans. Inform. Theory*, vol. IT-35, pp. 123–136, Jan. 1989.

[8] U. Madhow and M. L. Honig, "MMSE Interference Suppression for Direct-Sequence Spread-Spectrum CDMA," *IEEE Trans. Commun.*, vol. 42, pp. 3178–3188, Dec. 1994.

[9] A. Duel-Hallen, " Decorrelating Decision-Feedback Multiuser Detector for Synchronous Code-Division Multiple Access Channel ," *IEEE Trans. Commun.*, vol. COM-41, pp. 285–290, Feb. 1993.

[10] T. Guess and M. K. Varanasi, "Error Exponents for Maximum-Likelihood and Successive Decoders for the Gaussian CDMA Channel," *IEEE Trans. Inform. Theory*, vol. 46, pp. 1683–1691, Jul. 2000.

[11] M. K. Varanasi, "Decision Feedback Multiuser Detection: A Systematic Approach ," *IEEE Trans. Inform. Theory*, vol. IT-45, pp. 219–240, Jan. 1999.

[12] T. Guess and M. K. Varanasi, " Multiuser Decision-Feedback Receivers for the General Gaussian Multiple-Access Channel ," in *Proc. Thirty-fourth Allerton Conf. Commun., Contr. and Comput.*, Allerton, IL, pp. 190–199, Oct. 1996.

[13] R. A. Horn and C. R. Johnson, *Matrix Analysis.* Melbourne, Australia: Cambridge University Press, 1993.

Group Detectors for DS/CDMA Systems with Multipath Fading Channels

Stefano Buzzi *and* Marco Lops
Università degli Studi di Cassino
DAEIMI, {buzzi, lops}@unicas.it

Antonio Pauciullo
Università di Napoli "Federico II"
DIET, anpauciu@unina.it

Giuseppe Ricci
Università degli Studi di Lecce
Dip. di Ingegneria dell'Innovazione
giuseppe.ricci@unile.it

Abstract

Group detection is an appealing research topic, not only from a theoretical point of view, but also since its results can be directly applied to CDMA-based wireless cellular systems so as to account for out-of-cell interference and to multirate CDMA systems implemented with a multicode (MC) access technique. In this paper we present a new group detection strategy for asynchronous DS/CDMA Systems operating over frequency-selective fading channels. The proposed detection structure is a two stage-one: the first stage is a linear filter, aimed at suppressing the effect of the unwanted users signals, while the second stage is a non-linear block, implementing a maximum likelihood detection rule on the set of desired users signals. As to the linear stage, we consider both the zero-forcing (ZF) and the minimum mean square error (MMSE) approaches; in particular, based on the amount of prior knowledge on the interference parameters which is available to the receiver and on the affordable computational complexity, we come up with several receiving structures, which trade off system performance with complexity and prior information. Interestingly, all of the proposed structures are amenable to adaptive implementations, wherein only the parameters from the users to be decoded are assumed known. As to the performance assessment, simulation results confirm that the new receivers outperform previously derived detection structures.

1 Introduction

Given a Direct-Sequence Code Division Multiple Access (DS/CDMA) communication system with K active users, a group detector jointly demodulates the information bits stream from a certain subset, $\mathcal{G}$ say, of the K transmitting users. The concept of group detection was first introduced by Varanasi in [1], wherein, with reference to a synchronous CDMA system, new receiving structures were derived, based on the application of the generalized likelihood ratio test (GLRT) strategy. Interestingly, it was shown that, by partitioning the K active users in subsets of smaller cardinality, it is possible to achieve performance levels close to those of the optimum multiuser detector, but with smaller complexity. It was also pointed out that group detection provides a unifying approach to multiuser detection, in that both the optimum multiuser detector and some suboptimal single-user detectors may be regarded as special cases of the detectors derived in [1]. In [2], the said group detection structures are extended to the case that the channel introduces frequency-selective fading.

In the very recent past, group detection has become an attracting and intriguing research topic, not only under a theoretical point of view, but also for its applicability to the design of advanced wireless cellular CDMA-based networks. Indeed, in a multicell environment, a base-station is to jointly detect the bits from the users present in its own covered area, while not being interested in demodulating the signals originating from the out-of-cell users: in this context, group detectors appear much more suited than classical multiuser detectors, as they permit to take into account the presence of the (unwanted) inter-cell interference. Likewise, in a multirate CDMA network implemented with a multicode (MC) access technique [3], high data-rate users are assigned multiple signature waveforms, so that an high-rate bit stream is multiplexed onto multiple low-rate signals: in such a scenario, the mobile receiver of an high-rate user is thus interested in decoding multiple bits at each signaling interval, i.e. it is to behave like a group detector. Due to the strong relevance of these applications, many research efforts have been recently devoted to blind group detection, i.e. to the design of group detectors capable of being implemented with as limited as possible prior knowledge on the interference structure, but capable of achieving performance close to that of their non-blind counterparts. In this context, we cite here the work [4], wherein a semi-blind implementation of the detector in [1] has been presented, and the work [5], wherein linear group detectors are derived for an uplink CDMA channel, and the tools of subspace tracking theory are applied so as to come up with blind adaptive implementations. Finally, in [6, 7] a new two-stage group detection structure has been introduced for AWGN channels; interestingly, such a structure, which has been shown to outperform all of the previously derived group detectors, can be implemented in a blind adaptive fashion by resorting to both subspace tracking procedures and to the RLS algorithm.

Following on such a track, in this work we generalize the receivers in [6, 7] to the relevant case that the channel introduces frequency-selective fading. In particular, based on the amount of prior knowledge on the interfering signals parameters that is available to the receiver, and on the affordable receiver complexity, we come up with several detection structures that trade off performance with complexity and prior knowledge. Computer simulations show that the proposed receiving structures exhibit satisfactory performance and outperform the corresponding single-users detectors.

The paper is organized as follows. In the next Section we present the system model, while Section 3 is devoted to the synthesis of the proposed group detection structure, for both the cases of an uplink and downlink communication. In Section 4 computer simulation results are provided and discussed, while, finally, concluding remarks are drawn in Section 5.

2 System Model

Let us consider an asynchronous DS/CDMA network with K active users. The propagation channel is assumed to be a Rayleigh-distributed frequency-selective fading one, introducing L diversity paths: otherwise stated, the channel coherence bandwidth is smaller than the CDMA signals bandwidth, W say. We also assume that the channel coherence time is larger than the bit signaling time of the CDMA network. Therefore, the baseband equivalent of the useful signal from a given user consists of the superposition of L replicas of the transmitted signal, spaced by integer multiples of $1/W$ apart, each affected by a complex zero-mean Gaussian random gain. Accordingly, assuming a BPSK modulation format, the complex envelope of the received signal can be written as

$$r(t) = \sum_{k=0}^{K-1} \sum_{m=-P}^{P} \sum_{l=0}^{L-1} A_k b_k(m) \alpha_{k,l}(m) s_k(t - mT_b - \tau_k - l/W) + w(t) \tag{1}$$

In the above expression, A_k is the amplitude of the k-th user signal; $\alpha_{k,l}(m)$ is the k-th user, l-th replica channel fading realization in the m-th signaling interval; $\{b_k(m)\}_{m=-P}^{+P}$ represents the bit stream of the $k-$th user, modeled as a sequence of independent and identically distributed binary variates, each taking on values in the set $\{-1;1\}$; $T_b = NT_c$ is the bit interval of the network, with T_c the chip-interval and N the processing gain; $[\tau_0, \ldots, \tau_{K-1}]$ is a set of delays; $2P+1$ is the transmitted frame length; $s_k(t)$ is the *signature* assigned to the $k-$th user, expressed as

$$s_k(t) = \sum_{n=0}^{N-1} c_{kn} u_{T_c}(t - nT_c)$$

with $\{c_{kn}\}_{n=0}^{N-1}$ the $k-$th spreading code, and $u_{T_c}(\cdot)$ a unit-height rectangular pulse supported on the interval $[0, T_c]$; finally, $w(t)$ is the thermal noise term, modeled as a sample function from a complex white Gaussian process with Power Spectral Density (PSD) $2\mathcal{N}_0$. Notice that the model (1) is general enough to include most cases of relevant interest. Indeed, if we let $\alpha_{k,l}(m) = \alpha_{k',l}(m)\ \forall k, k' = 0, \ldots, K-1$, and $\tau_0 = \ldots, = \tau_{K-1}$, then the signal (1) models a downlink (i.e. base-to-mobile) cellular channel, wherein all of the signals propagate through the same channel and are synchronously transmitted. Conversely, if the above assumptions are not met, expression (1) represents the received signal in an uplink (i.e. mobile-to-base) cellular system, wherein all of the users asynchronously transmit their signals, which, in turn, propagate through different multipath channels. Likewise, if the fading coefficients $\alpha_{k,l}(m)$ are actually independent of the temporal index m, i.e. they remain constant over the whole packet duration, then the signal (1) does represent the received signal in a slowly varying fading channel, wherein the packet duration $(2P+1)T_b$ is smaller than the channel coherence time.

Now, let us assume that the users to be demodulated are the first G (with $G < K$), i.e. $\mathcal{G} = \{0, 1, \ldots, G-1\}$, and that $\tau_0 = \ldots = \tau_{G-1} = 0$, i.e. the users of interest transmit synchronously. This last assumption, which is naturally met in a multirate/multicode CDMA scenario, is made so as to simplify notation, but what follows can be easily extended also to the case that the system is fully asynchronous. In order to take a decision on the bits transmitted in the p-th signaling interval, $\boldsymbol{b}_G(p) = [b_0(p), \ldots, b_{G-1}(p)]^T$ (with $(\cdot)^T$ denoting transpose), we project the received waveform $r(t)$ along the unit vectors of the $NM-$dimensional orthonormal system:

$$\left\{ \frac{1}{\sqrt{T_{OS}}} u_{T_{OS}}(t - pT_b - iT_{OS}) \right\}_{i=0}^{NM-1} \tag{2}$$

with $T_{OS} = T_c/M$ and M a positive integer. Notice that, in defining the system (2) we have accounted for possible signal-space oversampling, in the sense specified in [8]. In particular, choosing $M > 1$, although resulting in some complexity increase, allows reducing the representation error due to the system asynchrony, and, eventually, achieving better performance. Projecting the waveform (1) onto the orthonormal set (2) yields the following NM–dimensional observables vector:

$$\boldsymbol{r}(p) = \sum_{k=0}^{G-1}\sum_{l=0}^{L-1} A_k b_k(p)\alpha_{k,l}(p)\boldsymbol{s}_{k,l}^{0} + \boldsymbol{z}(p) + \boldsymbol{w}(p) = \sum_{k=0}^{G-1} A_k b_k(p)\boldsymbol{S}_k\boldsymbol{\alpha}_k(p) + \boldsymbol{z}(p) + \boldsymbol{w}(p) \tag{3}$$

with

$$\boldsymbol{S}_k = \left[\boldsymbol{s}_{k,0}^{0}, \boldsymbol{s}_{k,1}^{0}, \ldots, \boldsymbol{s}_{k,L-1}^{0}\right] \qquad \boldsymbol{\alpha}_k(p) = \left[\alpha_{k,0}(p), \alpha_{k,1}(p), \ldots, \alpha_{k,L-1}(p)\right]^T \tag{4}$$

In equation (3) the first term on the right-hand-side (RHS) represents the contribution from the bits to be decoded, while $\boldsymbol{w}(p)$, the thermal noise discretized version, is a white complex Gaussian vector with covariance matrix $2\mathcal{N}_0\boldsymbol{I}_{NM}$[1]. Finally, in $\boldsymbol{z}(p)$ all of the remaining interfering signals have been lumped, i.e. the multiaccess interference (MAI) plus the intersymbol interference (ISI) induced by the channel dispersivity.

3 Detector Design

To begin with, first notice that, if no MAI were present at the receiver, if the fading coefficients $\alpha_{k,l}(p)$ were known to the receiver $\forall k = 0, \ldots, G-1$ and $\forall l = 0, \ldots, L-1$, and if the ISI were negligible[2], the minimum error-probability receiver would implement the following maximum likelihood (ML) decision rule:

$$\widehat{\boldsymbol{b}}_G(p) = \arg \min_{\boldsymbol{x}_G \in \{+1,-1\}^G} \left\| \boldsymbol{r}(p) - \sum_{k=1}^{G}\sum_{l=0}^{L-1} A_k x_k \alpha_{k,l}(p)\boldsymbol{s}_{k,l}^{0} \right\|^2 \tag{5}$$

with $\|\cdot\|$ denoting the Euclidean norm and x_k denoting the k-th entry of $\boldsymbol{x}_G$. Here, in order to account for the MAI plus ISI term $\boldsymbol{z}(p)$, we resort to a two-stage architecture: according to the scheme depicted in figure 1, we first feed the observables to a linear filter, represented by the $NM \times Q$-dimensional matrix $\boldsymbol{D}$ and aimed at suppressing MAI, and then forward its output to a non-linear block implementing a ML decision rule with respect to the G users of interest only, i.e.,

$$\widehat{\boldsymbol{b}}_G(p) = \arg \min_{\boldsymbol{x}_G \in \{+1,-1\}^G} \left[\left(\boldsymbol{r}(p) - \sum_{k=1}^{G} A_k x_k \boldsymbol{S}_k\boldsymbol{\alpha}_k(p)\right)^H \boldsymbol{D}(\boldsymbol{D}^H\boldsymbol{D})^{-1}\boldsymbol{D}^H \left(\boldsymbol{r}(p) - \sum_{k=1}^{G} A_k x_k \boldsymbol{S}_k\boldsymbol{\alpha}_k(p)\right)\right] \tag{6}$$

with $(\cdot)^H$ denoting conjugate transpose. Otherwise stated, we assume that the MAI contribution (i.e. the term $\boldsymbol{D}^H\boldsymbol{z}(p)$) at the output of the linear filter is vanishingly small, and implement the ML decision rule on the new observables *neglecting* the MAI contribution. It is worth pointing out that the second-stage of the receiver requires knowledge of the fading vectors $\boldsymbol{\alpha}_0(p), \ldots, \boldsymbol{\alpha}_{G-1}(p)$ affecting the signals from the users in $\mathcal{G}$. Here and in the following we assume that these vectors are known to the receiver or have been perfectly estimated; in practice, however, pilot symbols are needed in order to obtain estimates of the channel coefficients realizations: the impact of the estimation

---

[1] $\boldsymbol{I}_{NM}$ the identity matrix of order NM.

[2] This hypothesis is usually met in CDMA communication systems wherein $(L-1)/W$ is typically much smaller than T_b.

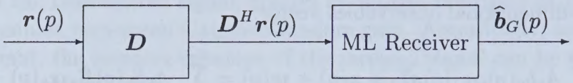

$\boldsymbol{r}(p)$
$\boldsymbol{D}$
$\boldsymbol{D}^H\boldsymbol{r}(p)$
ML Receiver
$\widehat{\boldsymbol{b}}_G(p)$

Figure 1: Block-scheme of the group detector.

errors on the system performance is not analyzed in this paper and is left for future work. Conversely, notice that no knowledge on the channel realizations for the interfering users is here required.

The next step is now the synthesis of the linear filter $\boldsymbol{D}$. In [6, 7], two possible design criteria for the matrix $\boldsymbol{D}$ have been proposed, which led to "group-versions" of the classical single-user decorrelating and MMSE receivers for an AWGN channel. Even though the received signal model (1) and its discrete-time counterpart (3) are general enough to include the already illustrated cases of relevant interest, the detector design necessarily requires splitting the derivation. In what follows, we thus extend the optimization criteria proposed in [7] so as to come up with detection structures for the downlink and uplink cellular frequency-selective fading channel.

3.1 The Downlink frequency-selective channel

As already anticipated, in a downlink (point-to-multipoint) communication scenario all of the signals propagate through the same channel, i.e. $\boldsymbol{\alpha}_0(p) = \ldots = \boldsymbol{\alpha}_{K-1}(p) = \boldsymbol{\alpha}(p)$, and, according to our previous assumptions, such a vector is known to the receiver. As a consequence, the signal (3) ends up similar to the received signal in a CDMA System with no fading distortion (i.e. in an AWGN channel). Accordingly, we propose that the matrix $\boldsymbol{D}$ is an $NM \times G$-dimensional matrix solving the following constrained minimization problem:

$$\begin{cases} E\left[\left\|\boldsymbol{D}^H(p)\boldsymbol{h}(p)\right\|^2 \Big| \boldsymbol{\alpha}(p)\right] = \min \\ \boldsymbol{D}^H(p)\boldsymbol{S}_G(p) = \boldsymbol{K} \end{cases} \tag{7}$$

wherein $\boldsymbol{S}_G(p) = [\boldsymbol{S}_0\boldsymbol{\alpha}_0(p), \ldots, \boldsymbol{S}_{G-1}\boldsymbol{\alpha}_{G-1}(p)]$ is an $NM \times G$-dimensional matrix containing on its columns the *effective* signatures from the users to be decoded, and $\boldsymbol{K}$ is a $G \times G$ matrix, which we assume to have full-rank. The problem (7) represents the extension to the group detection framework of the minimum mean-output-energy (MOE) constrained minimization problems considered in [9] with reference to single-user detection in an AWGN channel. It may be easily solved resorting to Lagrangian techniques. In particular, if we let $\boldsymbol{h}(p) = \boldsymbol{r}(p) - \boldsymbol{w}(p)$ we end up with a "group" decorrelating detector, while, on the contrary, if we let $\boldsymbol{h}(p) = \boldsymbol{r}(p)$, we obtain a "group" MMSE receiver. In both cases, the solution to (7) is written as:

$$\boldsymbol{D}(p) = \boldsymbol{R}^{+}_{\boldsymbol{hh}|\boldsymbol{\alpha}(p)}\boldsymbol{S}_G(p)\left(\boldsymbol{S}_G(p)^H\boldsymbol{R}^{+}_{\boldsymbol{hh}|\boldsymbol{\alpha}(p)}\boldsymbol{S}_G(p)\right)^{+}\boldsymbol{K}^H \tag{8}$$

wherein $(\cdot)^+$ denotes the Moore-Penrose generalized inverse, and

$$R_{hh|\alpha(p)} = E\left[h(p)h^H(p)|\alpha(p)\right]$$

is the conditional covariance matrix of the vector $h(p)$.

It may be easily shown that the solution (8) depends, through the fading vector $\alpha(p)$, on the temporal index p; accordingly, the filter $D(p)$ ends up time-invariant in the case of slow fading, i.e. if the vector $\alpha(p)$ remains constant over the entire frame duration. For not-slow fading, conversely, the filter $D(p)$ is actually time-varying, thus implying that, in the limit of fast fading, it is to be computed in each signaling interval. This might be a prohibitive computational effort, especially in mobile, low-power transceivers. In order to circumvent this drawback, we propose an alternative constrained minimization problem, i.e. we require that the filter D is an $NM \times G$-dimensional matrix solving the problem:

$$\begin{cases} E\left[\left\|D^H h(p)\right\|^2\right] = \min \\ D^H S_{LG} = K' \end{cases} \tag{9}$$

with $S_{LG} = [S_0, \ldots, S_{G-1}]$ and K' an $LG \times LG$-dimensional matrix which we again assume to have full-rank. If we let $h(p) = r(p) - w(p)$ or $h(p) = r(p)$, the solution to the problem (9) is written as

$$D = R_{hh}^+ S_{LG}\left(S_{LG}^H R_{hh}^+ S_{LG}\right)^+ K'^H \tag{10}$$

and, for Wide-Sense-Stationary (WSS) fading, is time-invariant. In the above equation, the matrix $R_{hh} = E[h(p)h^H(p)]$ is the unconditional covariance matrix of the vector $h(p)$. The following observations can be now made on the found solutions:

1. The filter D in (10) offers poorer performance than the filter $D(p)$ in (8), in that it relies on a smaller amount of prior information (i.e. it can be realized based on the knowledge of only the second-order statistics of the observables and does not require knowledge of the fading vector realizations) and is to fulfill a larger number of constraints (i.e. $(LG)^2$ instead of G^2 scalar constraints). Nonetheless, the filter (10) has the appealing property of being time-invariant, and may be useful for implementation in fast fading environments, wherein limited computational resources prevent an *on-line* computation of the time-varying filter (8).

2. Due to the optimum combining rule (6) it is easily seen that the choice of the full-rank constraint matrices K and K' does not affect the decision rule, i.e. any full-rank matrix leads to the same receiver.

3. If we let $h(p) = r(p) - w(p)$ and none of the columns of the matrix $S_G(p)$ lies in the range span of the matrix $R_{hh|\alpha(p)}$, the filter $D(p)$ in (8) behaves like a Zero-Forcing (ZF) receiver, namely it is capable of totally nullifying the external intererference, thus implying that the decision rule (6) is the optimal processing for the new observables (i.e., $D(p)^H r(p)$). The same statement can be made for the filter (10) if none of the columns of the matrix S_{LG} lies in the range span of the matrix R_{hh}.

4. If the conditions specified in the previous item are fulfilled, and if we consider a synchronous system with no oversampling, the proposed two-stage receiving structure with the filter D given by (8) ends up coincident with one of the group detectors derived by Varanasi in [2].

3.2 The Uplink frequency-selective channel

In an uplink (multipoint-to-point) communication channel, the transmitted signals propagate asynchronously and through different channels. Assuming, in keeping with the previous derivations, that the receiver has knowledge of the fading vectors for the users in $\mathcal{G}$, we propose that the filter $\boldsymbol{D}$ is an $NM \times LG$-dimensional matrix solving the problem:

$$\begin{cases} E\left[\left\|\boldsymbol{D}^H(p)\boldsymbol{h}(p)\right\|^2 \middle| \boldsymbol{\alpha}_0(p), \ldots, \boldsymbol{\alpha}_{G-1}(p)\right] = \min \\ \boldsymbol{D}^H(p)\boldsymbol{S}_G(p) = \boldsymbol{K} \end{cases} \tag{11}$$

Letting, again, $\boldsymbol{h}(p) = \boldsymbol{r}(p) - \boldsymbol{w}(p)$ or $\boldsymbol{h}(p) = \boldsymbol{r}(p)$, the solution can be compactly written as:

$$\boldsymbol{D}(p) = \boldsymbol{R}^+_{\boldsymbol{hh}|\boldsymbol{\alpha}_0(p),\ldots,\boldsymbol{\alpha}_{G-1}(p)}\boldsymbol{S}_G(p)\left(\boldsymbol{S}_G(p)^H\boldsymbol{R}^+_{\boldsymbol{hh}|\boldsymbol{\alpha}_0(p),\ldots,\boldsymbol{\alpha}_{G-1}(p)}\boldsymbol{S}_G(p)\right)^+\boldsymbol{K}^H \tag{12}$$

with, in turn, $\boldsymbol{R}_{\boldsymbol{hh}|\boldsymbol{\alpha}_0(p),\ldots,\boldsymbol{\alpha}_{G-1}(p)} = E[\boldsymbol{h}(p)\boldsymbol{h}^H(p)|\boldsymbol{\alpha}_0(p), \ldots, \boldsymbol{\alpha}_{G-1}(p)]$. Like in the downlink framework, we thus obtain a time-varying filter which ends up time-invariant in a slow fading environment. In order to come up with a receiving structure with lower computational complexity, namely with a time-invariant filter $\boldsymbol{D}$, we can also resort to the filter (10), solving the problem (9) and that can be applied in an uplink communication channel as well. The following observations can be now made:

1. Solution (12) achieves improved performance with respect to solution (10). This fact may be intuitively justified through the same arguments exposed in the previous subsection.
2. The full-rank matrix $\boldsymbol{K}$ in (11) is actually irrelevant.
3. If we let $\boldsymbol{h}(p) = \boldsymbol{r}(p) - \boldsymbol{w}(p)$ and none of the columns of the matrix $\boldsymbol{S}_G(p)$ lies in the range span of the matrix $\boldsymbol{R}_{\boldsymbol{hh}|\boldsymbol{\alpha}_0(p),\ldots,\boldsymbol{\alpha}_{G-1}(p)}$, the filter $\boldsymbol{D}(p)$ in (12) behaves like a Zero-Forcing (ZF) receiver, namely it is capable of totally nullifying the external intererference, thus implying that the decision rule (6) is the optimal processing for the new observables (i.e., $\boldsymbol{D}(p)^H\boldsymbol{r}(p)$).

As a final remark, notice that the detection structures (8) may be implemented in a blind adaptive fashion by properly modifying the procedures reported in [7]. Due to lack of space, we do not dwell here on such an issue.

4 Simulation Results

In order to test the effectiveness of the proposed group detectors, we have considered a DS/CDMA system employing Gold codes with processing gain $N = 31$ and with perfect power control (i.e., $A_0 = \ldots = A_{K-1}$). First of all we focus on a downlink communication affected by slow fading and with the multipath diversity order $L = 3$. In figure 2 we have considered a synchronous systems with $K = 15$ users and $G = 8$, and we have represented the average Bit-Error-Rate for the users in $\mathcal{G}$ versus the average received energy contrast per bit γ_0. In particular the performance of the group decorrelating and MMSE receivers, labeled as GDEC and GMMSE, respectively, and with the filter $\boldsymbol{D}$ given by (8), is contrasted with that of the classical single-user linear decorrelating and

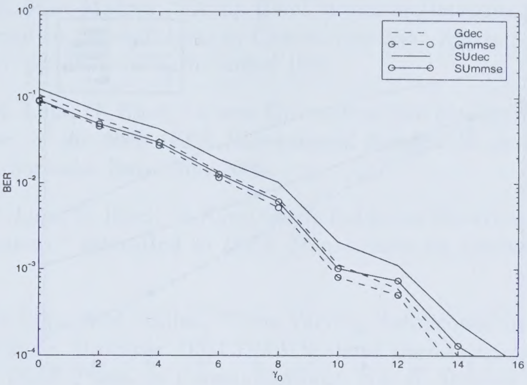


Figure 2: Bit-Error-Rate of the proposed detectors versus received energy contrast per bit in a slow fading downlink channel with $K = 15$, $G = 8$ and $M = 1$.

MMSE receivers. It is seen that the group detectors achieve a performance improvement with respect to the classical single-user detectors.

In figure 3, instead, we focus on an asynchronous trasmission in an uplink channel with fast fading; we represent the error probability for the newly proposed group detectors with the filter $\boldsymbol{D}$ given by (10) versus that of the single-user detectors. Again, we have $L = 3$, and we have let $K = 10$, $G = 8$ and $M = 1$, i.e. we did not resort to oversampling. The plots have been obtained through a Montecarlo counting procedure, and are the result of an average over a set of 40 random realizations of the MAI delays. It is seen that in such a severe scenario the adoption of group detectors offers a remarkable performance improvement with respect to the single-user detectors. From both figures 2 and 3 it is also seen that the performance gap between the GMMSE and the GDEC receiver is negligible. We point out that this is not a general trend and indeed there are instances (not shown here for the sake of brevity) in which the GMMSE considerably outperforms the GDEC receiver. Finally, other simulations have revealed that, in keeping with the results in [7, 8], in heavily loaded asynchronous systems even moderate values of the oversampling factor M yield noticeable performance improvements.

5 Concluding Remarks

In this paper we have considered the problem of group detection for DS/CDMA Systems operating over frequency-selective fading channels. Starting upon the results presented in [6, 7], wherein an AWGN propagation channel had been considered, we have presented new group detection structures able to operate over frequency-selective fading channels. The proposed detectors have a two-stage structure: the first stage is a linear filter rejecting the interfering users signals and aimed at ensuring the system immunity to the near-far effect; the second stage, instead, is a non-linear device implementing a maximum-likelihood detection strategy with respect to the bits to be decoded. The said

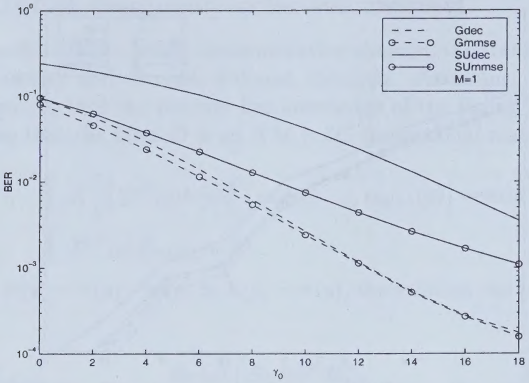


Figure 3: Bit-Error-Rate of the proposed detectors versus received energy contrast per bit in a fast fading uplink channel with $K = 10$, $G = 8$, and $M = 1$.

structure has been specialized to the cases that either the fading remains constant for many signaling intervals or it changes at every bit-interval, as well as to the cases that all of the signals propagate either synchronously through the same channel (downlink scenario) or asynchronously through different channels, as happens in an uplink cellular communication. The simulation results have confirmed that the proposed detection structures exhibit satisfactory performance and outperform the corresponding single-user structures.

Further work on this topic is devoted to the design of adaptive implementations of the said receivers, wherein no prior knowledge is assumed on the interfering users signatures waveforms, and on the fading coefficients affecting the signals to be decoded.

References

[1] M.K. Varanasi, "Group Detection for Synchronous Gaussian Code-Division Multiple-Access Channels," *IEEE Trans. on Information Theory*, Vol. 41, pp. 1083-1096, July 1995.

[2] M. K. Varanasi, "Parallel Group Detection for Synchronous CDMA Communication Over Frequency-Selective Rayleigh Fading Channels," *IEEE Trans. on Information Theory*, Vol. 42, pp. 116-128, January 1996.

[3] U. Mitra, "Comparison of Maximum Likelihood-based Detection for two Multi-rate Access Schemes for CDMA Signals," *IEEE Trans. on Communications*, Vol. 47, pp. 64-77, January 1999.

[4] G. Ricci, M. K. Varanasi, "Blind Multiuser Detection of L Out Of K Active Users Of an N Users Synchronous CDMA System," *IEEE Vehicular Technology Conference*, Houston, Texas, May 16-19, 1999.

[5] X. Wang, A. Host-Madsen, "Group-Blind Multiuser Detection for Uplink CDMA," *IEEE Journal on Selected Areas in Communications: Wireless Communication Series*, Vol. 17, pp. 1971-1984, November 1999.

[6] S. Buzzi, M. Lops, G. Ricci, "A new Group detection strategy for DS/CDMA Systems," *Proc. of the 2000 IEEE International Symposium on Information Theory (ISIT '00)*, Sorrento, Italy, June 2000.

[7] S. Buzzi, M. Lops, G. Ricci, "A Novel Group Detection Structure for Direct-Sequence CDMA Systems," submitted to *IEEE Transactions on Communication*, February 2000.

[8] S. Buzzi, M. Lops, A.M. Tulino, "Time-Varying Narrowband Interference Rejection in Asynchronous Multiuser DS/CDMA Systems over Frequency-Selective Fading Channels," *IEEE Trans. on Communications*, Vol. 47, October 1999.

[9] X. Wang, H.V. Poor, "Blind Multiuser Detection: A Subspace Approach," *IEEE Trans. on Information Theory*, Vol. 44, pp. 677-690, March 1998.

A Family of Likelihood Ascent Search Detectors Achieving Local Maximum Likelihood with an Arbitrary Neighborhood Size for CDMA Multiuser Detection[1]

Yi Sun
Department of Electrical Engineering
The City College of City University of New York
New York, NY 10031
E-mail: yisun@ee-mail.engr.ccny.cuny.edu

Abstract – **In our previous work we demonstrated that there exists a set of local maximum likelihood (LML) detectors defined with neighborhood size one. In this paper, the concept of the LML detection is extended to an arbitrary neighborhood size, thus the previously studied LML detectors as well as the well-known optimum detector are special instances. It is shown that every LML detector is associated with a local minimum error probability defined with the corresponding neighborhood size. As the neighborhood size is one, two, etc., up to the total number of users, the computational complexity of the LML detector is linear, quadratic, etc., up to exponential in the total number of users. Then, we propose a family of local-maximum-likelihood likelihood-ascent-search (LMLAS) detectors with an arbitrary neighborhood size. Every LMLAS detector is shown to be an LML detector. An LMLAS detector monotonically increases likelihood step by step, and thus converges to an LML point in a finite number of search steps with probability one. Following any detector, an LMLAS detector can reduce the error probability of the initial detector to a local minimum or not change it when the initial detector is an LML point with the same or greater neighborhood size with probability one.**

I. INTRODUCTION

Verdú [2][3] showed that an optimum multiuser detector in terms of maximum likelihood can achieve significant performance improvement over the conventional detector. However, its computational complexity grows exponentially with the number of active users. Many low-complexity suboptimal multiuser detectors [1] have been developed, including the suboptimal tree-type maximum-likelihood sequence detectors, linear suboptimum detectors, successive interference cancellation (SIC) detector, decorrelating decision-feedback detector (DDFD), and parallel interference cancellation (PIC) detector.

A family of local maximum likelihood (LAS) multiuser detectors was developed recently [4][5], which originated from a family of modified Hopfield neural network algorithms [6].The family of LAS detectors has many provable good characteristics [4][5]. All LAS detectors guarantee likelihood ascent from one search step to another, thus converging to a fixed point in a finite number of search steps with probability one. A LAS detector improves another detector by reducing its error probability unless the given initial detector is a fixed point of the LAS detector with probability one. In particular, it is shown that in the family of LAS detector, there is a set of wide-sense sequential LAS (WSLAS) detectors that with

---

[1] This work was in part prepared through collaborative participation in the Advanced Telecommunications / Information Distribution Research Program (ATIRP) Consortium sponsored by the U.S. Army Research Laboratory under Cooperative Agreement DAAL01-96-2-0002, and supported in part by a grant #69708-00-29 from The City University of New York PSC-CUNY Research Award Program.

probability one achieve local maximum likelihood (LML) detection defined with neighborhood size one. Thus, they achieve the local minimum error probabilities with the same neighborhood size.

The WSLAS detectors achieve the LML detection with the smallest neighborhood consisting of bit vectors differing from the demodulated vector by no more than one bit. In contrast, the optimum detector which achieves the global maximum likelihood (GML) detection achieves the LML detection with the largest neighborhood consisting of all possible bit vectors. In this paper, the concept of LML detection is extended to *any* neighborhood size. Then a family of local-maximum-likelihood likelihood-ascent-search (LMLAS) detectors is proposed.

II. LOCAL-MAXIMUM-LIKELIHOOD LIKELIHOOD-ASCENT-SEARCH DETECTORS

A. Received baseband CDMA signal

Consider a bit-synchronous Gaussian CDMA channel of K users with processing gain M. After sampled at the chip rate, the received one-shot CDMA signal in matrix form can be

$$\mathbf{y} = \mathbf{SAb} + \mathbf{n} \tag{1}$$

where the ith user transmits bit b_i spread by signature waveform $\mathbf{s}_i \in \Re^M$, $\|\mathbf{s}_i\| = 1$, with signal amplitude A_i. $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_K)$ is the signature waveform matrix of K users and $\mathbf{A} = \mathrm{diag}(A_1, \ldots, A_K)$. $\mathbf{n} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$ is an M-dimensional Gaussian random vector with zero mean and covariance matrix $\sigma^2\mathbf{I}$. $\mathbf{y}$ is an M-dimensional random vector of sufficient statistic of $\mathbf{b}$. The transpose of the signature matrix $\mathbf{S}^{\mathrm{T}}$ is equivalent to a matched-filter bank. By applying $\mathbf{S}^{\mathrm{T}}$ to the data of (1), the output of the matched-filter bank is obtained by

$$\mathbf{r} = \mathbf{S}^{\mathrm{T}}\mathbf{y} = \mathbf{RAb} + \mathbf{z} \tag{2}$$

where $\mathbf{R} = \mathbf{S}^{\mathrm{T}}\mathbf{S}$ is the crosscorrelation matrix of signatures with unit diagonal elements $R_{ii} = 1$, $\mathbf{z} = \mathbf{S}^T\mathbf{n}$ is a Gaussian random vector with zero mean and covariance matrix $\sigma^2\mathbf{R}$.

B. Likelihood function and metric

Throughout this paper, we assume that all users independently transmit bits −1 and 1 with equal probability. From (2), a likelihood function can be written as

$$\Lambda(\mathbf{r} \mid \mathbf{b}) = \frac{1}{(2\pi\sigma^2)^{K/2} \mid \mathbf{R} \mid^{1/2}} \exp\left(-\frac{(\mathbf{RAb} - \mathbf{r})^T \mathbf{R}^{-1} (\mathbf{RAb} - \mathbf{r})}{2\sigma^2} \right) \tag{3}$$

where if the inverse of $\mathbf{R}$ does not exist, $\mathbf{R}^{-1}$ denotes its pseudo-inverse and $|\mathbf{R}|$ is the product of nonzero eigenvalues of $\mathbf{R}$.

It is well-known that the minimization of the error probability leads to the optimum (GML) detector that selects the hypothesis $\mathbf{b}^{\mathrm{GML}}$ maximizing the likelihood function globally, i.e.

$$\mathbf{b}^{\mathrm{GML}} = \arg \max_{\mathbf{b} \in \{-1,1\}^K} \Lambda(\mathbf{r} \mid \mathbf{b}),\ \forall \mathbf{r} \in \Re^K. \tag{4}$$

The error probability of the GML detector is

$$P_e(\mathbf{b}^{\mathrm{GML}}) = 1 - \frac{1}{2^K} \int_{\Re^K} \max_{\mathbf{b} \in \{-1,1\}^K} \Lambda(\mathbf{r} \mid \mathbf{b}) dr\,. \tag{5}$$

From (3), a metric can be defined by

$$f(\mathbf{r} \mid \mathbf{b}) = \frac{1}{2}(\mathbf{RAb} - \mathbf{r})^T \mathbf{R}^{-1} (\mathbf{RAb} - \mathbf{r}) = \frac{1}{2}\mathbf{b}^T\mathbf{W}\mathbf{b} - \mathbf{q}^T\mathbf{b} + \frac{1}{2}\mathbf{r}^T\mathbf{R}^{-1}\mathbf{r} \tag{6}$$

where $\mathbf{W} = \mathbf{ARA}$ and $\mathbf{q} = \mathbf{Ar}$. Since $R_{kk} = 1$, $W_{kk} = A_k^2$ for $\forall k$.

For $\forall \mathbf{a}, \mathbf{b} \in \{-1, 1\}^K$, if and only if $\Lambda(\mathbf{r} \mid \mathbf{a}) < \Lambda(\mathbf{r} \mid \mathbf{b})$, then $f(\mathbf{r} \mid \mathbf{a}) > f(\mathbf{r} \mid \mathbf{b})$. We can rewrite (4) as

$$\mathbf{b}^{\mathrm{GML}} = \arg \min_{\mathbf{b}\in\{-1,1\}^K} f(\mathbf{r}\mid\mathbf{b}),\ \forall \mathbf{r}\in \Re^K. \tag{7}$$

The negative gradient of metric $f(\mathbf{r} \mid \mathbf{b})$ with respect to $\mathbf{b}$ is

$$\mathbf{h}(\mathbf{b}) \equiv -\nabla f(\mathbf{b}) = -\mathbf{W}\mathbf{b} + \mathbf{q}. \tag{8}$$

C. A generalized LMLAS detector

A generalized LMLAS (GLMLAS) detector is proposed as follows.

GLMLAS detector: Given an initial $\mathbf{b}(0) \in \{-1, 1\}^K$ and a sequence of bit index subsets $L(t) \subseteq \{1, ..., K\}$ such that $1 \le |L(t)| \le J$ for all $t \ge 0$. For all $t \ge 0$, repeat the following update until termination. If

$$\sum_{i\in L(t)} b_i(t)\left(h_i(t) + \sum_{j\in L(t)} W_{ij} b_j(t)\right) < 0 \tag{9}$$

where $h_i(t)$ is the ith component of the negative metric gradient $\mathbf{h}(t)$ evaluated at $\mathbf{b}(t)$, then the bits are updated by

$$b_k(t+1) = \begin{cases} -b_k(t), & \text{if } k \in L(t), \\ b_k(t), & \text{otherwise,} \end{cases} \tag{10}$$

and the negative metric gradient is updated by

$$\mathbf{h}(t+1) = \mathbf{h}(t) + 2\sum_{k\in L(t)} b_k(t)\mathbf{W}_k \tag{11}$$

where $\mathbf{W}_k$ is the kth column vector of matrix $\mathbf{W}$. If $\mathbf{b}(t) = \mathbf{b}^f$ for $0 \le t_1^f \le t \le t_2^f$ and

$$\bigcup_{t_1^f \le t \le t_2^f} \left\{ \mathbf{b}^f - 2\sum_{i\in L(t)} b_i^f \mathbf{e}_i \right\} = N_J(\mathbf{b}^f) \setminus \{\mathbf{b}^f\} \tag{12}$$

where $N_J(\mathbf{b}^f)$ denotes the neighborhood of $\mathbf{b}^f$ with neighborhood size J (which will be defined in next section), then terminate with $\mathbf{b}^f$ of the finally demodulated bit vector. □

The integer J in the GLMLAS detector is called neighborhood size because as we will see, the GLMLAS detector converges to an LML point with neighborhood size J. $L(t)$ determines the vector to be compared with the current vector $\mathbf{b}(t)$. If the flip condition (9) is satisfied, $\mathbf{b}(t)$'s bits whose indices belong to $L(t)$ are flipped and a new vector $\mathbf{b}(t+1)$ is accepted. Specifying a neighborhood size J and an index subset sequence $L(t)$ for $t \ge 0$ in the GLMLAS detector, one produces an LMLAS detector. Hence, the GLMLAS detector defines a family of LMLAS detectors.

D. An example of LMLAS detector

Given below is an example how to specify the sequence of index subsets.

E-LMLAS detector: Denote $\Omega(l) \equiv \{L \subseteq \{1,\ldots,K\} \mid |L| = l\}$. (i) Given an initial $\mathbf{b}(0) \in \{-1, 1\}^K$. (ii) For $j = J$ to 1, for each $L \in \Omega(j)$, update bits according to (9)-(11). (iii) If no flip occurs in step (ii), terminate; otherwise, go to (ii). □

Clearly, the E-LMLAS detector belongs to the family of the LMLAS detectors. Performance of the E-LMLAS detector is examined in simulation.

E. Computational complexity of update and required amount of memory

Consider $\forall L \subseteq \{1, \ldots, K\}$. Due to the symmetry of $\mathbf{R}$, condition (9) can be rewritten as

$$\sum_{i\in L} b_i h_i + \sum_{i\in L} W_{ii} + \sum_{i\in L}\sum_{j\in L, j>i} 2W_{ij} b_i b_j < 0, \tag{13}$$

where $2W_{ij}$ can be stored as one number. Note that the multiplication of a number with a bit is this number with its sign changed if and only if the bit is -1. Suppose that the change of sign is not counted in computational complexity. The check of condition (13) needs a total of $|L| + |L| + |L|(|L| - 1)/2 - 1 = (|L|^2 + 3|L| - 2)/2$ additions, which is the number of additions for the update when condition (9) is not satisfied. If condition (9) is satisfied, the extra number of additions for updating **h** in (11) with an arbitrary crosscorrelation matrix is $K|L|$. In this case, the total number of additions for the update is $[|L|^2 + (2K+3)|L| - 2]/2$. If there are zero crosscorrelations, the update of **h** needs fewer additions. Finally, we note that the update of **h** is involved with vector additions. By parallel computing, the time of a vector addition is the same as the time of a scalar addition. Thus, parallel computing can significantly reduce the computational time in an LMLAS detector especially when K is much larger than J. In particular, this is also true for LAS detectors [4][5] which were originally motivated by implementation on a Hopfield neural network hardware for fast parallel massive computing.

In addition to storage of $\mathbf{W}$ and $\mathbf{q}$, an LMLAS detector needs only to store $\mathbf{h}(t)$ and $\mathbf{b}(t)$, each of which has a size of K no matter how large the neighborhood size is.

III. LOCAL MAXIMUM LIKELIHOOD POINTS

A. Local maximum likelihood point

In this section, we consider LML points defined with an arbitrary neighborhood size.

Definition 1: For $\forall \mathbf{b} \in \{-1, 1\}^K$, the neighborhood of $\mathbf{b}$ with size J is defined as the set

$$N_J(\mathbf{b}) = \{\mathbf{a} \in \{-1, 1\}^K \mid \|\mathbf{a} - \mathbf{b}\|_1 \le J\} \tag{14}$$

where $\|\mathbf{a} - \mathbf{b}\|_1$ denotes the Hamming distance between $\mathbf{a}$ and $\mathbf{b}$. □

For $\forall \mathbf{a} \in N_J(\mathbf{b})$, $\mathbf{a}$ and $\mathbf{b}$ differ by at most J bits. For $\forall \mathbf{b} \in \{-1, 1\}^K$, the total number of bit vectors in $N_J(\mathbf{b})$ is $|N_J(\mathbf{b})| = \sum_{j=0}^{J}\binom{K}{j}$; in particular, $|N_K(\mathbf{b})| = 2^K$.

Definition 2: For $\forall \mathbf{r} \in \Re^K$, consider $\forall \mathbf{b} \in \{-1, 1\}^K$ and its neighborhood $N_J(\mathbf{b})$. If $\Lambda(\mathbf{r} \mid \mathbf{b}) \ge \Lambda(\mathbf{r} \mid \mathbf{a})$ for $\forall \mathbf{a} \in N_J(\mathbf{b})\backslash\{\mathbf{b}\}$, then $\mathbf{b}$ is a local maximum likelihood (LML) point with neighborhood size J. □

Definition 3: For $\forall \mathbf{r} \in \Re^K$, the set of LML points with neighborhood size J is defined by $\Psi_J^{\mathrm{LML}}(\mathbf{r}) = \{\mathbf{b} \in \{-1, 1\}^K \mid \Lambda(\mathbf{r} \mid \mathbf{b}) \ge \Lambda(\mathbf{r} \mid \mathbf{a}),\ \forall \mathbf{a} \in N_J(\mathbf{b})\backslash\{\mathbf{b}\}\}$. □

Property 1: For $\forall \mathbf{r} \in \Re^K$, $\Psi_1^{\mathrm{LML}}(\mathbf{r}) \supseteq \Psi_2^{\mathrm{LML}}(\mathbf{r}) \supseteq \ldots \supseteq \Psi_K^{\mathrm{LML}}(\mathbf{r})$. □

Property 2: For $\forall \mathbf{a}, \mathbf{b} \in \Psi_J^{\mathrm{LML}}(\mathbf{r})$, if $\mathbf{a} \ne \mathbf{b}$, then $\|\mathbf{a} - \mathbf{b}\|_1 \ge J + 1$. In particular, $\|\mathbf{a} - \mathbf{b}^{\mathrm{GML}}\|_1 \ge J + 1$. □

Lemma 1: For $\forall \mathbf{r} \in \Re^K$, $\mathbf{b} \in \Psi_J^{\mathrm{LML}}(\mathbf{r})$ iff

$$\sum_{i \in L} A_i b_i \left(r_i - \sum_{j \notin L} R_{ij} A_j b_j \right) \ge 0 \tag{15}$$

for all $L \subseteq \{1,\ldots,K\}$ such that $1 \le |L| \le J$. □

According to Lemma 1, for $\forall \mathbf{b} \in \Psi_J^{\mathrm{LML}}(\mathbf{r})$, the total number of inequalities in (15) that $\mathbf{b}$ satisfies is equal to $\sum_{j=1}^{J}\binom{K}{j}$. In particular, for $\mathbf{b}^{\mathrm{GML}}$ the total number is $2^K - 1$. This suggests the relative computational complexity for the search of an LML point with a given neighborhood size J. In other words, if the neighborhood size is one, two, etc., up to K, the computational complexity is linear, quadratic, etc., up to exponential in the total number of users. The GML point is specified by the most conditions. That is, for $\forall \mathbf{r} \in \Re^K$, $\mathbf{b}^{\mathrm{GML}}$ is the GML point if and only if

$$\sum_{i \in L} A_i b_i^{\mathrm{GML}} \left(r_i - \sum_{j \notin L} R_{ij} A_j b_j^{\mathrm{GML}} \right) \ge 0 \tag{16}$$

for all $L \subseteq \{1,\ldots,K\}$ such that $1 \le |L| \le K$.

B. Local maximum likelihood region

Limit points of a search detector are finally demodulated vectors. The observation region of a limit point is the decision region of the detector with an according initial.

Definition 4: Let $\Psi^{\phi}(\mathbf{r}) \subseteq \{-1, 1\}^K$ be the set of limit points of a search detector ϕ with observation $\mathbf{r}$. The limit region of a limit point $\mathbf{b}$ in the observation space is defined as $V^{\phi}(\mathbf{b}) = \{\mathbf{r} \in \Re^K \mid \mathbf{b} \in \Psi^{\phi}(\mathbf{r})\}$. □

Definition 5: For $\forall \mathbf{b} \in \{-1, 1\}^K$, the LML region of $\mathbf{b}$ in the observation space is defined as $V_J^{\text{LML}}(\mathbf{b}) = \{\mathbf{r} \in \Re^K \mid \mathbf{b} \in \Psi_J^{\text{LML}}(\mathbf{r})\}$. □

Property 3: For $\forall \mathbf{b} \in \{-1, 1\}^K$, the LML region of $\mathbf{b}$ with neighborhood size J is $V_J^{\text{LML}}(\mathbf{b}) = \{\mathbf{r} \in \Re^K \mid \mathbf{r}$ satisfies (15)$\}$. □

Property 4: For $\forall \mathbf{b} \in \{-1,1\}^K$, $V_1^{\text{LML}}(\mathbf{b}) \supseteq V_2^{\text{LML}}(\mathbf{b}) \supseteq \ldots \supseteq V_K^{\text{LML}}(\mathbf{b})$. □

Property 5: For $\forall \mathbf{b} \in \{-1, 1\}^K$, $V_J^{\text{LML}}(\mathbf{b})$ is convex. □

As the neighborhood size increases, the LML region becomes more and more complicated with more and more hyperplanes, and closer and closer to the decision region $V_K^{\text{LML}}(\mathbf{b})$ of the GML detector.

C. LML detector and local minimum error probability

To better understand the family of LMLAS detectors, we introduce the concepts of LML detector and local minimum error probability.

Definition 6: For a detector $\mathbf{b}^{\phi}$, if $\mathbf{b}^{\phi}(\mathbf{r}) \in \Psi_J^{\text{LML}}(\mathbf{r})$ for $\forall \mathbf{r} \in \Re^K$, $\mathbf{b}^{\phi}$ is an LML detector with neighborhood size J and denoted by $\mathbf{b}^{\phi} \in \Psi_J^{\text{LML}}$. □

In a notation similar to (4) for the GML detector, an LML detector with neighborhood size J can be written as

$$\mathbf{b}_J^{\text{LML}} = \arg \max_{\mathbf{b} \in N_J(\mathbf{b}_J^{\text{LML}})} \Lambda(\mathbf{r} \mid \mathbf{b}), \forall \mathbf{r} \in \Re^K. \tag{17}$$

When $J = K$, (17) becomes (4) because the largest neighborhood for any vector is $\{-1, 1\}^K$.

Property 6: If $\mathbf{b}^{\phi} \in \Psi_J^{\text{LML}}$, then $\mathbf{b}^{\phi} \in \Psi_I^{\text{LML}}$ for $\forall I < J$; or conversely, if $\mathbf{b}^{\phi} \notin \Psi_I^{\text{LML}}$, then $\mathbf{b}^{\phi} \notin \Psi_J^{\text{LML}}$ for $\forall J > I$. □

Property 6 indicates that an LML detector with neighborhood size J is also an LML detector with neighborhood size smaller than J. The GML detector is a special instance of LML detector that is defined with the largest neighborhood size K and so is an LML detector of any neighborhood size.

Definition 7: Consider two detectors $\mathbf{b}^{\phi}$ and $\mathbf{b}^{\varphi}$. If $\mathbf{b}^{\phi}(\mathbf{r}) \in N_J[\mathbf{b}^{\varphi}(\mathbf{r})]$ for $\forall \mathbf{r} \in \Re^K$, then $\mathbf{b}^{\phi}$ is said to be in the neighborhood of $\mathbf{b}^{\varphi}$ and denoted by $\mathbf{b}^{\phi} \in N_J(\mathbf{b}^{\varphi})$. □

For any detector $\mathbf{b}^{\phi}$ with demodulated vector $\mathbf{b}^{\phi}(\mathbf{r})$ for $\forall \mathbf{r} \in \Re^K$, an error probability $P_e(\mathbf{b}^{\phi})$ is associated. Thus, $P_e(\mathbf{b}^{\phi})$ is a function of $\mathbf{b}^{\phi}$ and is determined by $\mathbf{b}^{\phi}(\mathbf{r}) \in \{-1, 1\}^K$ for all $\mathbf{r} \in \Re^K$. Because of the geometric relationship between two detectors each of which is in other's neighborhood, we can define a local minimum error probability as follows.

Definition 8: If $P_e(\mathbf{b}^{\phi}) \le P_e(\mathbf{b}^{\varphi})$ for $\forall \mathbf{b}^{\varphi} \in N_J(\mathbf{b}^{\phi})$, then $P_e(\mathbf{b}^{\phi})$ is a local minimum error probability with neighborhood size J. □

Property 7: $\mathbf{b}^{\phi}$ achieves a local minimum error probability with neighborhood size J iff $\Pr\{\mathbf{b}^{\phi} \in \Psi_J^{\text{LML}}\} = 1$. □

Corollary 1: Every LML detector with neighborhood size J achieves a local minimum error probability with the same neighborhood size J. $\square$

Theorem 1: Let $\mathbf{b}_J^{\text{LML}}$ and $\mathbf{b}_I^{\text{LML}}$ denote LML detectors with neighborhood sizes J and I for $J < I$, respectively. If $\Pr\{\mathbf{b}_J^{\text{LML}} \in N_I(\mathbf{b}_I^{\text{LML}})\} = 1$, then $P_e(\mathbf{b}_J^{\text{LML}}) \geq P_e(\mathbf{b}_I^{\text{LML}})$ where the equality holds iff $\Pr\{\Lambda(\mathbf{b}_J^{\text{LML}}) = \Lambda(\mathbf{b}_I^{\text{LML}}) \mid \mathbf{b}_J^{\text{LML}} \in N_I(\mathbf{b}_I^{\text{LML}})\} = 1$. $\square$

In Theorem 1, the condition $\Pr\{\Lambda(\mathbf{b}_J^{\text{LML}}) = \Lambda(\mathbf{b}_I^{\text{LML}}) \mid \mathbf{b}_J^{\text{LML}} \in N_I(\mathbf{b}_I^{\text{LML}})\} = 1$ is hardly true in general, and thus $P_e(\mathbf{b}_J^{\text{LML}}) > P_e(\mathbf{b}_I^{\text{LML}})$ if $\Pr\{\mathbf{b}_J^{\text{LML}} \in N_I(\mathbf{b}_I^{\text{LML}})\} = 1$. In other words, if an LML detector is in the neighborhood of another LML detector with a larger neighborhood size with probability one, the former has a higher error probability.

IV. PROPERTIES OF LMLAS DETECTORS

A. Likelihood ascent and stability

The GLMLAS detector has the following property from which several other properties can be derived.

Theorem 2: For $\forall \mathbf{r} \in \Re^K$, the GLMLAS detector guarantees monotonic likelihood ascent,

$$\Lambda[\mathbf{r} \mid \mathbf{b}(t+1)] \geq \Lambda[\mathbf{r} \mid \mathbf{b}(t)], \qquad \forall t \geq 0, \tag{18}$$

where the equality holds iff $\mathbf{b}(t+1) = \mathbf{b}(t)$. $\square$

Corollary 2: For $\forall \mathbf{b}(0) \in \{-1, 1\}^K$ and $\mathbf{r} \in \Re^K$, let $\mathbf{b}^{\text{GLMLAS}}$ be a fixed point of the GLMLAS detector with initial $\mathbf{b}(0)$. (i) If $\mathbf{b}(0) \notin \Psi^{\text{GLMLAS}}(\mathbf{r})$, then $\mathbf{b}^{\text{GLMLAS}} \neq \mathbf{b}(0)$ and $\Lambda(\mathbf{r} \mid \mathbf{b}^{\text{GLMLAS}}) > \Lambda[\mathbf{r} \mid \mathbf{b}(0)]$; (ii) if $\mathbf{b}(0) \in \Psi^{\text{GLMLAS}}(\mathbf{r})$, then $\mathbf{b}^{\text{GLMLAS}} = \mathbf{b}(0)$. $\square$

As indicated by Theorem 2, the family of LMLAS detectors are indeed *likelihood ascent search* (LAS) detectors.

Theorem 3: (i) For $\forall \mathbf{r} \in \Re^K$, after a finite number of search steps, the GLMLAS detector converges to a fixed point; (ii) the GLMLAS detector converges to a fixed point in a finite number of steps with probability one. $\square$

Theorem 4: For $\forall \mathbf{r} \in \Re^K$, let $\Im^{\text{GLMLAS}}(\mathbf{r})$ and $\Psi^{\text{GLMLAS}}(\mathbf{r})$ be the limit point set and the fixed point set of the GLMLAS detector, respectively. Then $\Im^{\text{GLMLAS}}(\mathbf{r}) = \Psi^{\text{GLMLAS}}(\mathbf{r})$. $\square$

B. Fixed point set of LMLAS detectors

The following theorem shows that the LMLAS detectors are indeed *local maximum likelihood* (LML) detectors.

Theorem 5: Let $\Psi_J^{\text{GLMLAS}}(\mathbf{r})$ be the fixed point set of GLMLAS detector with neighborhood size J. Then $\Psi_J^{\text{GLMLAS}}(\mathbf{r}) = \Psi_J^{\text{LML}}(\mathbf{r})$, $\forall \mathbf{r} \in \Re^K$. $\square$

Corollary 3: $\Pr(\Psi_J^{\text{GLMLAS}} = \Psi_J^{\text{LML}}) = 1$. $\square$

Corollary 4: For any initial $\mathbf{b}(0) \in \{-1, 1\}^K$, $\mathbf{b}_J^{\text{GLMLAS}} \in \Psi_J^{\text{LML}}$. $\square$

As indicated by Corollary 4, every LMLAS detector with neighborhood size J is an LML detector with the same neighborhood size. Because of this, all characteristics of an LML point and its LML region are characteristics of the fixed point of an LMLAS detector.

Property 8: For $\forall \mathbf{r} \in \Re^K$, $\Psi_1^{\text{GLMLAS}}(\mathbf{r}) \supseteq \Psi_2^{\text{GLMLAS}}(\mathbf{r}) \supseteq \ldots \supseteq \Psi_K^{\text{GLMLAS}}(\mathbf{r})$. $\square$

C. Monotonic descent of error probability

The monotonic likelihood ascent of the GLMLAS detector directly results in the monotonic descent of error probability.

Theorem 6: The GLMLAS detector monotonically reduces error probability,

$$P_e[\mathbf{b}(t+1)] \leq P_e[\mathbf{b}(t)], \qquad \forall t \geq 0 \tag{19}$$

where the equality holds if and only if $\Pr[\mathbf{b}(t+1) = \mathbf{b}(t)] = 1$. □

Theorem 7: Let $\mathbf{b}_J^{\text{LMLAS}}$ be an LMLAS detector with an arbitrary neighborhood size J with initial $\mathbf{b}(0)$. Then

$$P_e(\mathbf{b}_J^{\text{LMLAS}}) \le P_e[\mathbf{b}(0)] \tag{20}$$

where $P_e(\mathbf{b}_J^{\text{LMLAS}})$ is a local minimum error probability with neighborhood size J and the equality holds iff $\Pr[\mathbf{b}(0) \in \Psi_J^{\text{LML}}] = 1$. □

Theorem 7 characterizes one important property of the family of LMLAS detectors. The initial $\mathbf{b}(0)$ of the LMLAS detector can be any detector of any kind. A local minimum error probability achieved by an LML detector may not be necessarily smaller than an error probability achieved by a non-LML detector. However, as we see from Theorem 7, unless a non-LML detector differs from an LML detector only on observations of zero probability measure, its error probability can be reduced by a followed LMLAS detector to a local minimum error probability with the same neighborhood size of the LMLAS detector.

In a simulation test, the WSLAS detector reduces the bit error rates of the conventional detector, decorrelator, MMSE detector, and PIC detector. Since the WSLAS detector is an LML detector with neighborhood size one, this means that none of these well-known detectors is an LML detector with neighborhood size one. Due to Property 6, none of these detectors is an LML detector with any neighborhood size in the conditions of the simulation.

Theorem 7 is applicable to the case when two LMLAS detectors are cascaded and leads to the following corollary. Due to Theorem 5 and Property 8, an LMLAS detector can not improve any LMLAS (or LML) detector with the same or a larger neighborhood size. In particular, no LMLAS detector can improve the GML detector. The following corollary indicates that an LMLAS detector with a larger neighborhood size can improve an LMLAS detector with a smaller neighborhood size unless the later is also an LML detector with the larger neighborhood size with probability one. Note that an LMLAS detector with a smaller neighborhood size can be an LML detector with a larger neighborhood size with probability one, say in the case when its initial is an LML point with the larger neighborhood size with probability one.

Corollary 5: Let $\mathbf{b}_J^{\text{LMLAS}}$ be a fixed point of an LMLAS detector with neighborhood size J with initial $\mathbf{b}_I^{\text{LMLAS}}$, a fixed point to which another LMLAS detector with neighborhood size I converges. If $J > I$, then $P_e(\mathbf{b}_J^{\text{LMLAS}}) \le P_e(\mathbf{b}_I^{\text{LMLAS}})$ where the equality holds iff $\Pr(\mathbf{b}_I^{\text{LMLAS}}(\mathbf{r}) \in \Psi_J^{\text{LML}}(\mathbf{r})) = 1$. □

V. CONCLUSIONS

The GML detector has been well studied in literature and the LML detectors with neighborhood size one was studied by our recent work. In this paper, we extend the concept of local maximum likelihood detection to an arbitrary neighborhood size, and propose a family of LMLAS detectors.

The inequalities of an LML point with an arbitrary neighborhood size are given. The number of the inequalities are equal to the number of vectors in a neighborhood, which suggests the computational complexity of an LML detector with the same neighborhood size. When the neighborhood size is equal to one, two, etc., up to the total number of users, the complexity is linear, quadratic, etc., up to exponential in the number of users, respectively.

The inequalities also define the set of LML points with a given observation, and define the LML region in the observation space for a given bit vector. As the neighborhood size increases, the number of elements in an LML point set decreases and the difference between an LML region and the optimum (GML) decision region reduces while the complexity for decision of an LML point increases.

The LMLAS detectors guarantee monotonic likelihood ascent in search, thus guaranteeing convergence to a fixed point in a finite number of search steps with probability one. Any fixed point of an LMLAS detector with a given neighborhood size is an LML point with the same neighborhood size, and vice versa. An LMLAS detector with a given neighborhood size monotonically reduces error probability step by step and finally reaches a local minimum error probability defined with the same neighborhood size. The error probability of any detector can be reduced to a local minimum by a followed LMLAS detector unless with probability one this detector itself is an LML detector with the same or greater neighborhood size. As an LML detector, an LMLAS detector achieves a lower error probability with a higher computational complexity as the neighborhood size is larger. One can make a tradeoff between error performance and computational complexity by choosing a neighborhood size in an LMLAS detector. The BER floor exists for then GLMLAS detector that does not achieve the GML detection as SNR increases. However, the BER floor is at a lower level, starts to appear at a higher SNR, and finally disappears as the neighborhood size of the LMLAS detector becomes larger and finally becomes the total number of users. When the neighborhood size is small, an LMLAS detector can be benefited from a moderate near-far situation. This near-far resistance for the LMLAS detector with a small neighborhood size alleviates the demand on power control.

This work can be extended to signal detection problems where signals to be detected have multiple levels such as digital image restoration and reconstruction or complex symbols such as symbol detection in single user channels as well as OFDM systems.

REFERENCES

[1] S. Verdú, *Multiuser detection*, Cambridge University Press, New York, 1998.

[2] S. Verdú, "Minimum probability of error for asynchronous Gaussian multiple-access channels," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 85-96, Jan. 1986.

[3] S. Verdú, "Optimum, multiuser asymptotic efficiency," *IEEE Trans. Commun*, vol. COM-34, pp. 890-897, Sept. 1986.

[4] Y. Sun, “A family of likelihood ascent search detectors for CDMA multiuser detection,” submitted to *IEEE Trans. on Information Theory*.

[5] Y. Sun, "A family of linear complexity likelihood ascent search detectors for CDMA multiuser detection," in *Proc. IEEE 6th Int. Symp. on Spread-Spectrum Tech. & Appli.*, ISSSTA'2000, Parsippany, NJ, Sept. 6-8, 2000.

[6] Y. Sun, “A generalized updating rule for modified Hopfield neural network for quadratic optimization,” *Neurocomputing*, 19 (1998), pp. 133-143.

[7] M. K. Varanasi and B. Aazhang, "Multistage detection in asynchronous code-division multiple-access systems," *IEEE Trans. Commun.*, vol. 38, Apr. 1990, pp. 509-519.

[8] M. K. Varanasi and B. Aazhang, "Near-optimum detection in synchronous code-division multiple access systems," *IEEE Trans. Commun.*, vol. 39, May 1991, pp. 725-736.

[9] H. L. Van Trees, *Detection, Estimation, and Modulation Theory – Part I Detection, Estimation, and Linear Modulation Theory*, John Wiley & Sons, New York, 1968.

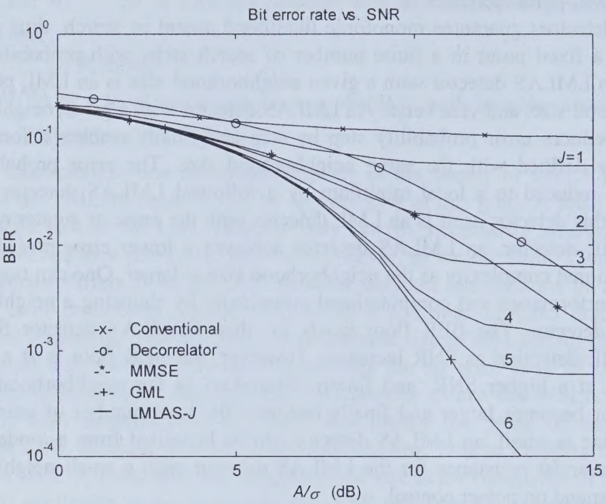


Fig. 1. BER versus SNR. Processing gain is $M = 15$; number of users is $K = 10$; system is perfectly power-controlled. The initial is the conventional detector. The BER of the LMLAS-J detector with $J = 6$ is overlapped with the BER of the GML detector.

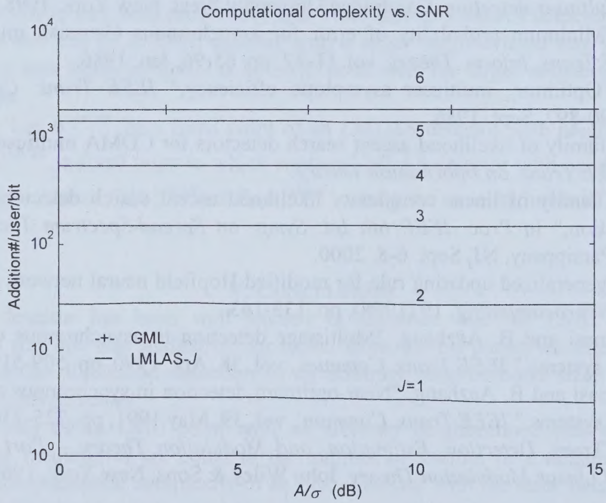


Fig. 2. Average number of additions per user per bit versus SNR. The condition is the same as in Fig. 1. The computational complexity is approximately independent of SNR.

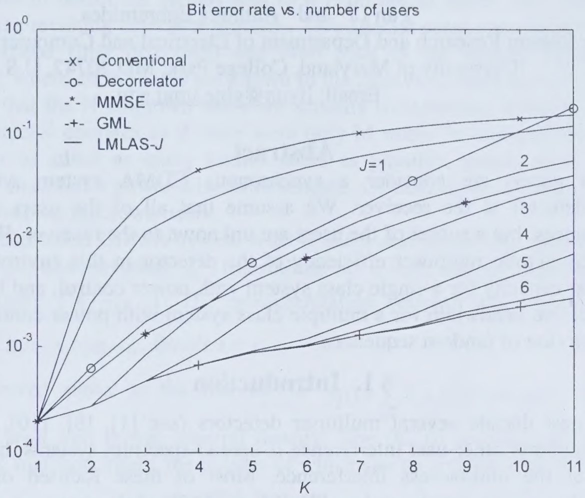


Fig. 3. BER versus number of users. Processing gain is $M = 11$, SNR = 11 dB, and the system is perfectly power-controlled.

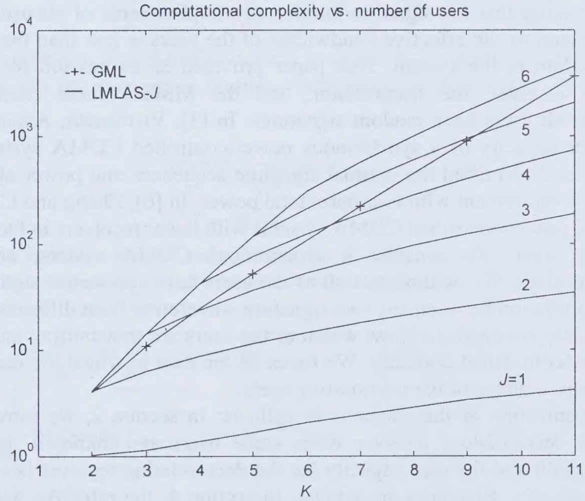


Fig. 4. Average number of additions per user per bit versus number of users. The condition is the same as in Fig. 3. The computation of the GML detector is optimized.

Multiuser Detectors for Incompletely Known Symmetric Signals

Yun Li and Anthony Ephremides
Institute for System Research and Department of Electrical and Computer Engineering,
University of Maryland, College Park, MD 20742, U.S.A.
Email: liyun@glue.umd.edu

Abstract

In this paper, we consider a synchronous CDMA system with a multiuser decorrelating detector at the receiver. We assume that all of the users have symmetric signature sequences, but a subset of the users are unknown to the receiver. We first calculate the SIR and asymptotic multiuser efficiency of the detector in this environment. We then identify the user capacity for a single class system with power control, and the user capacity region and effective bandwidth for a multiple class system with power control. The result is compared to the case of random sequences.

1. Introduction

In the past decade several multiuser detectors (see [1], [8], [10], and [11]) were analyzed that suppress other-user interference in spread-spectrum systems, by making use of the structure of the multiaccess interference. Most of these focused on the multiuser efficiency and the near far resistance (see [7], [9], and [12]) of the receivers. Since improving network capacity is always the aim for both the physical layer and network layer design ([4]), more attention is now being paid to the receiver performance in power-controlled CDMA systems and to the resulting user capacity (see [2], [3], [5], and [6]).

In [2], Tse and Hanly derived an "effective bandwidth" characterization of the user capacity in the sense that the signal-to-interference requirements of all users can be met if and only if the sum of the effective bandwidths of the users is less than the total number of degrees of freedom in the system. That paper provided an expression for the case of the matched filter receiver, the decorrelator, and the MMSE linear receiver, under the assumption that all users have random signatures. In [3], Vismanath, Anantharam, and Tse studied the user capacity of a synchronous power-controlled CDMA system with MMSE linear receiver, and identified the optimal signature sequences and power allocation scheme to meet the SIR requirement with minimum total power. In [6], Zhang and Chong studied the user capacity of power-controlled CDMA systems with linear receivers in fading channels.

In this paper, we consider a synchronous CDMA system, equipped with a decorrelating receiver. We assume that all of the users have symmetric signature sequences, i.e., the crosscorrelation between any two signature waveforms from different users are same. The multiuser detector needs to know which of the users are transmitting and their signature sequences in order to detect correctly. We focus on the case in which the receiver knows the signatures of only a subset of the transmitting users.

The organization of this paper is as follows: In section 2, we introduce the signal model and the decorrelating receiver when some users are unknown. In section 3, the effective bandwidth and the user capacity for the decorrelating receiver is computed for the case in which all user sequences are known. In section 4, the effective bandwidth and the user capacity are derived when some of the signal sequences are unknown. In section 5, we close with a summary and conclusions.

2. Signal Model and the (M,N) Decorrelator

Assume there are K users sending message to the receiver and the receiver only knows M of them ($1 \le M \le K$), the remaining N ($N = K - M$) users are unknown, in the sense that the receiver does not know of their presence (not does it know their signatures). The receiver tries to detect the signals sent to it from the M users. The notation (M, N) is used to denote the resulting detection problem. Two special cases correspond to M=1 (that is single user matched filter detection) and N=0 (that is the regular multiuser detectors discussed in [1], where all the users are known to the receiver.) Since in our case the receiver does not know that the N unknown users are actually transmitting, it has to treat them like noise, so the detector operates as if there were only M users. In a power-controlled system, the system tries to admit as many known users as possible, which meet a specific SIR requirement by means of a power allocation scheme. Of course, the system can not exercise power control over the unknown users.

The synchronized CDMA system with K users and processing gain L is as follows: User *i* (i=1,2,…,K) has amplitude A_i , transmitted bit $b_i = \pm 1$, and signature waveform $S_i(t)$, $0 \le t \le T$, satisfying $\int_0^T S_i^2(t)dt = 1$. The crosscorrelation between any two of these users is given by: $\rho_{ij} = \int_0^T S_i(t)\cdot S_j(t)dt = \rho$, for all $i \neq j$.

The received signal at the receiver is: $y(t) = \sum_{i=1}^{K} A_i b_i S_i(t) + \sigma \cdot n(t)$, with n(t) white Gaussian noise with unit power spectral density. The (K, 0) receiver passes the received signal through K matched filters, the outputs of which can be written in matrix form as:

$$\mathbf{y} = \mathbf{R}\cdot\mathbf{A}\cdot\mathbf{b} + \mathbf{n} \tag{1}$$

with

$$\mathbf{y} = \begin{bmatrix} \int_0^T y(t)\cdot S_1(t)dt \\ \int_0^T y(t)\cdot S_2(t)dt \\ \vdots \\ \int_0^T y(t)\cdot S_K(t)dt \end{bmatrix}, \mathbf{n} = \begin{bmatrix} \int_0^T n(t)\cdot S_1(t)dt \\ \int_0^T n(t)\cdot S_2(t)dt \\ \vdots \\ \int_0^T n(t)\cdot S_K(t)dt \end{bmatrix}, \mathbf{R} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}, \mathbf{A} = \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_K \end{bmatrix}, \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_K \end{bmatrix}.$$

Then the quantity $\mathbf{R}^{1}\cdot\mathbf{y}=\mathbf{A}\cdot\mathbf{b}+\mathbf{R}^{1}\cdot\mathbf{n}$ is used to detect **b**, by setting $\hat{b}_i = \mathrm{sgn}((\mathbf{R}^{-1}\mathbf{y})_i)$. The error probability and the asymptotic multiuser efficiency in detecting b_i for (K, 0) receiver are given by:

$$P_{\mathrm{dec},i}^{(K,0)}(\sigma) = Q\left(\frac{A_i}{\sigma\sqrt{(\mathbf{R}^{-1})_{ii}}}\right), \tag{2}$$

$$\eta_{\mathrm{dec},i}^{(K,0)} = (\mathbf{R}^{-1})_{ii} \; . \tag{3}$$

When the receiver knows only M of the K users, the detector operates actually exactly in the same way as the (M, 0) decorrelator, as shown in Figure 1. The outputs of the M matched filters are:

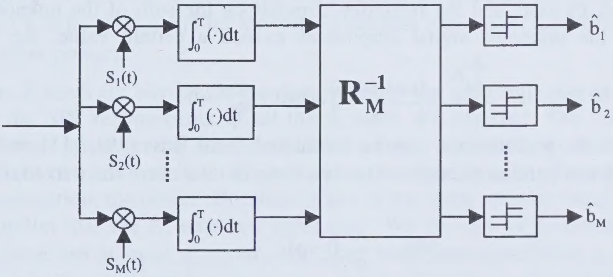


Figure 1: Structure of (M, 0) or (M, N) decorrelator

$$y_i = A_i b_i + \sum_{\substack{j=1 \\ j \neq i}}^{K} \rho_{ij} A_j b_j + n_i\,, \quad i = 1,2,\ldots,M \cdot \tag{4}$$

In vector form: $$\mathbf{y_M} = [\mathbf{R_M} \quad \Delta] \cdot \begin{bmatrix} \mathbf{A_M b_M} \\ \mathbf{A_N b_N} \end{bmatrix} + \mathbf{n_M} = \mathbf{R_M A_M b_M} + \Delta \mathbf{A_N b_N} + \mathbf{n_M}\,. \tag{5}$$

Where $\mathbf{R_M}$ is the $M \times M$ crosscorrelation matrix of the M known users; Δ is the $M \times N$ crosscorrelation matrix between the N unknown users and the M known users with $\Delta_{ij} = \rho$ for all $1 \le i \le M, 1 \le j \le N$; $\mathbf{A_M b_M}$ is $M \times 1$ vector with $(\mathbf{A_M b_M})_i = A_i b_i, i = 1,2,\cdots M$; and $\mathbf{A_N b_N}$ is $N \times 1$ vector with $(\mathbf{A_N b_N})_i = A_{M+i} b_{M+i}, i = 1,2,\cdots N$.

So, $$\mathbf{R_M^{-1}} \cdot \mathbf{y_M} = \mathbf{A_M b_M} + \mathbf{R_M^{-1}} \Delta \mathbf{A_N b_N} + \mathbf{R_M^{-1} n_M}\,, \tag{6}$$

And $$\hat{b}_i = \operatorname{sgn}((\mathbf{R_M^{-1}} \bar{\mathbf{y}}_{\mathbf{M}})_i) = \operatorname{sgn}(A_i b_i + (\mathbf{R_M^{-1}} \Delta \mathbf{A_N b_N})_i + (\mathbf{R_M^{-1} n_M})_i). \tag{7}$$

Here $(\mathbf{R_M^{-1} n_M})_i$ is the Gaussian noise with variance $(\mathbf{R_M^{-1}})_{ii}$. Comparing to the (M, 0) case, we see that $\mathbf{R_M^{-1}} \Delta \mathbf{A_N b_N}$ represents the extra interference term from the N unknown users. After some matrix calculation, we obtain:

$$(\mathbf{R_M^{-1}} \Delta \mathbf{A_N b_N})_i = \frac{\rho}{1 + \rho(M-1)} \sum_{j=M+1}^{K} A_j b_j,\ i = 1,2,\cdots M\,, \tag{8}$$

$$\text{and } (\mathbf{R_M^{-1}})_{ii} = \frac{1 + \rho(M-2)}{(1 + \rho(M-1))(1-\rho)}, \quad i = 1,2,\ldots,M. \tag{9}$$

Therefore the SIR of user *i* is given by:

$$\mathrm{SIR}_i = \frac{A_i^{\,2}}{\dfrac{\sigma^2}{1-\rho} \cdot \dfrac{1+\rho(M-2)}{1+\rho(M-1)} + \left[\dfrac{\rho}{1+\rho(M-1)}\right]^2 \displaystyle\sum_{j=M+1}^{K} A_j^2}\,, \tag{10}$$

and the error probability of user *i* is expressed as:

$$P_{\mathrm{dec},i}^{(M,N)}(\sigma) = 2^{-N} \sum_{b_{M+1} = \{+1,-1\}} \cdots \sum_{b_K = \{+1,-1\}} Q\left(\frac{A_i + \dfrac{\rho}{1+\rho(M-1)} \displaystyle\sum_{j=M+1}^{K} A_j b_j}{\sqrt{\dfrac{\sigma^2}{1-\rho} \dfrac{1+\rho(M-2)}{1+\rho(M-1)}}} \right). \tag{11}$$

Notice that when N=0, Eq. (11) is the same as Eq. (1), as expected. When M=1, Eq. (11) reduces to the matched filter case. From the error probability, expression of the asymptotic multiuser efficiency is calculated as:

$$\eta_{\mathrm{dec},i}^{(M,N)} = \max{}^2 \left\{ 0, \left(1 - \frac{\rho}{1+\rho(M-1)} \cdot \sum_{j=M+1}^{K} \frac{A_j}{A_i} \right) \cdot \sqrt{\frac{[1+\rho(M-1)](1-\rho)}{1+\rho(M-2)}} \right\}. \tag{12}$$

From Eq. (12), it is seen that the multiuser efficiency of the (M, N) case is smaller than that of the (M, 0) case, and the reduction depends on the sum of the unknown signal amplitude. When the unknown signal amplitudes exceed a certain value, the multiuser efficiency reduces to zero. If $\dfrac{\sum_{j=M+1}^{K} A_j}{A_i} \ge \left| \dfrac{1+\rho(M-1)}{\rho} \right|$, then $\eta_{\mathrm{dec},i}^{(M,N)} = 0$.

The asymptotic performance can be calculated from Eqs. (10), (11) and (12), as $L \to \infty, K \to \infty$, $\frac{K}{L} = \alpha$, and as the ratio of known users to total users stays fixed, i.e. $\frac{M}{K} = \beta$. We find that,

$$\mathrm{SIR}_i \to \frac{A_i^2}{\sigma^2}(1-\rho), \tag{13}$$

$$\lim_{K\to\infty} \eta_{\mathrm{dec},i}^{(M,N)} = \lim_{K\to\infty} (1-\rho) \max{}^2 \left\{ 0,\ 1 - \frac{\rho}{1+\rho(M-1)} \sum_{j=M+1}^{K} \frac{A_j}{A_i} \right\} = (1-\rho) \max{}^2 \left\{ 0, 1 - \frac{\rho N \bar{a}_i}{1+\rho(M-1)} \right\}. \tag{14}$$

Here the quantity $\bar{a}_i = \lim_{K\to\infty} \frac{1}{N} \sum_{j=M+1}^{K} \frac{A_j}{A_i}$ is the average relative energy of the unknown users with respects to user *i*. Notice that the interference from the unknown users vanishes asymptotically because it is weighted by the $\left[\frac{\rho}{1+(M-1)\rho}\right]^2$, which goes to 0 according to M^{-2}. Also notice that if $\rho \to 0$ the SIR is the same as that of the single user matched filter. This is because the signals become approximately orthogonal as $\rho \to 0$.

Assume for simplicity that $\bar{a}_i = 1$. We observe that

If $0<\rho<1$ is a constant, $\lim_{K\to\infty} \eta_{\mathrm{dec},i}^{(M,N)} = (1-\rho)\max{}^2\left\{0, 2-\frac{1}{\beta}\right\}$, (15)

If $\rho = \frac{1}{\sqrt{L}} \to 0$, $\lim_{K\to\infty} \eta_{\mathrm{dec},i}^{(M,N)} = \max{}^2\left\{0, 2-\frac{1}{\beta}\right\}$, and (16)

If $\rho = \frac{1}{L} \to 0$, $\lim_{K\to\infty} \eta_{\mathrm{dec},i}^{(M,N)} = \lim_{K\to\infty}\left(1-\frac{N}{L+M}\right)^2 = \left(\frac{1+2\alpha\beta-\alpha}{1+\alpha\beta}\right)^2$. (17)

3. User Capacity and Effective Bandwidth for (K,0) Decorrelator

Recall Eq. (10) for M=K; it yields

$$\mathrm{SIR}_i = \frac{A_i^2}{\sigma^2 \left(R_K^{-1}\right)_{ii}} = \frac{P_i}{\frac{\sigma^2}{1-\rho}\left[\frac{1+\rho(K-2)}{1+\rho(K-1)}\right]}. \tag{18}$$

From [2], in the expression of the SIR for a large system which is given by

$$\mathrm{SIR}_i = \frac{P_i}{\sigma^2 + \frac{1}{L}\sum_{\substack{j=1 \\ j\neq i}}^{K} I_{ji}(P_j, P_i, \mathrm{SIR}_i)}. \tag{19}$$

The quantity $I_{ji}(P_j, P_i, \mathrm{SIR}_i)$ is the effective interference that user *j* imposes on user *i*. Comparing Eq. (18) to Eq. (19), we can obtain the effective interference from each of the K-1 users to user *i* as:

$$I_{ji} = \frac{\sigma^2\rho^2 L}{[1+\rho(K-1)](1-\rho)} = \frac{P_i}{\mathrm{SIR}_i}\frac{\rho^2 L}{1+\rho(K-2)}, \quad j \neq i. \tag{20}$$

The effective interference from user *j* is independent of the power of user *j*. This property follows from the structure of the decorrelator. The decorrelator uses the projection of **y** onto $\{s_1, s_2, \cdots, s_{i-1}, s_{i+1}, \cdots, s_K\}^{\perp}$ to detect b_i, so the performance in detecting user *i* has nothing to do with the power of the other users. In [2], the effective interference in the decorrelator is given by P_i/β_i for a large system and for the case of random sequences, (which is also independent of the interfere power).

The K users are admissible into the system if one can find a power allocation scheme such that the SIR requirements of all the K users are satisfied. The maximal admissible number of users per degree of freedom without a power constraint is called the user capacity. Here we have two constraints to the maximal admissible number of users. One is from the power consumption; the power allocation to any of the users must be finite. The other is from our assumption that the K users are symmetric. We assume an L-dimensional space, and therefore there are at most γL symmetric users with crosscorrelation ρ, (for finite L and $0 \le \rho \le 1$, γ is 1. See section 5 for detail.) As stated in [2], The notion of effective bandwidth can be defined such that the SIR requirements of all the users can be met if and only if the sum of the effective bandwidths of the users is less than the total number of degrees of freedom in the system.

Suppose there is no power constraint, and that all users have the same SIR requirement β, i.e. $\mathrm{SIR}_i \geq \beta, i = 1,2,\cdots,K$. Since all the users have symmetric sequences and have same SIR requirement, their power levels should be the same. The obvious power allocation is then

$$P_i = \frac{\beta\sigma^2}{1-\rho}\left[\frac{1+\rho(K-2)}{1+\rho(K-1)}\right].\ i = 1,2,\cdots K\cdot \tag{21}$$

The allocated power increases as the number of user increases. Since $P_i < \frac{\beta\sigma^2}{1-\rho} < \infty$, there is no constraint imposed on the number of users from the point of view of the allocated power. However, the assumption of the existence of K symmetric users with crosscorrelation ρ requires that $K \leq \gamma L$. Therefore,

$$C_{\mathrm{dec}}^{(K,0)}(\beta) = \gamma \text{ users / degree of freedom.} \tag{22}$$

For finite L and $0 \leq \rho \leq 1$, the user capacity is 1 user per degree of freedom. This result is the same as for the case of random sequences considered in [2]. A natural question that arises is that, if $\gamma = 1$, the number of users is always less than or equal to the processing gain, so, why not let $\rho = 0$. In this way the same user capacity is achieved and less power is consumed. The answer is that in assigning sequences (and, hence, the value of ρ), the designer deals with a number of subscribers much greater than L. In general, fewer than L users are simultaneously present in the system at any given moments. So the case we are considering corresponds to a subset of a large number of users who can not have orthogonal sequences in L dimensions. Why should they be symmetric then? They need not be. We assume they are here, both for simplifying the analysis and for obtaining a benchmark reference point in the evaluation of the performance of this detector.

Now we add a power constraint, namely that $P_i \leq \overline{P}, i = 1,2,\cdots,K$. If there is only one user in the system, the power required to satisfy its SIR requirement is $\beta\sigma^2$. We therefore assume that $\overline{P} \geq \beta\sigma^2$.

If $\overline{P} \geq \frac{\beta\sigma^2}{1-\rho}$, then $P_i < \frac{\beta\sigma^2}{1-\rho} \leq \overline{P}$ is satisfied automatically, so the power constraint has no effect on the maximum number of user supportable in the system. We still have $K \leq \gamma L$, and the maximum number of users per degree of freedom that can be supported is γ. If $\beta\sigma^2 \leq \overline{P} < \frac{\beta\sigma^2}{1-\rho}$, it follows from $P_i \leq \overline{P}$ that $K \leq 1 - \frac{1}{\rho} + \frac{\beta\sigma^2}{\beta\sigma^2 - \overline{P}(1-\rho)}$. Therefore the maximum number of users per degree of freedom that can be supported is given by:

$$\min\left\{\frac{1}{L} - \frac{1}{\rho L} + \frac{\beta\sigma^2}{\beta\sigma^2 L - \overline{P}(1-\rho)L},\ \gamma\right\}\cdot \tag{23}$$

Now suppose that there are J classes, with user populations respectively given by $K_j, j = 1,2,\cdots J$, $(\sum_{j=1}^{J} K_j = K)$ and with specific SIR requirements $\mathrm{SIR}_j \geq \beta_j, j = 1,2,\cdots J$. Define $\overline{\beta} = \max_j \beta_j$. The power allocation is the same for all users in the same class, that is

$$P_j = \frac{\beta_j\sigma^2}{1-\rho}\left[\frac{1+\rho(K-2)}{1+\rho(K-1)}\right].\ j = 1,2,\cdots J\cdot \tag{24}$$

If there is no power constraint, since the maximum possible power is $\frac{\overline{\beta}\sigma^2}{1-\rho} < \infty$, there is no constraint on the number of users from the power point of view, and the user capacity constraint is $\sum_{j=1}^{J} K_j \le \gamma L$. From this, we can identify the effective bandwidth of the *j*th class user to be

$$e_{\mathrm{dec}}^{(K,0)}(\beta) = \frac{1}{\gamma} \text{ degree of freedom / user.} \tag{25}$$

For finite L and $0 \le \rho \le 1$, we have $\gamma=1$ and, hence, the effective bandwidth is 1 degree of freedom per user. This result is the same as for the random sequences, case considered in [2].

If we have a power constraint $P_j \le \bar{P}_j < \infty$, $j = 1,2,\cdots,J$, we first partition the J classes into three sets as follows:

$$J_1 = \left\{ j : \bar{P}_j \ge \frac{\beta_j \sigma^2}{1-\rho} \right\},\ J_2 = \left\{ j : \sigma^2 \beta_j \le \bar{P}_j < \frac{\beta_j \sigma^2}{1-\rho} \right\}, \text{ and } J_3 = \left\{ j : \bar{P}_j < \sigma^2 \beta_j \right\}.$$

For class $j \in J_3$, it is not possible to satisfy the power constraint, so $\sum_{j \in J_3} K_j = 0$.

If J_2 is not a null set, then $K \le 1 - \frac{1}{\rho} + \frac{\beta_j \sigma^2}{\beta_j \sigma^2 - \bar{P}_j (1-\rho)}$ for any $j \in J_2$. Therefore the user capacity constraint can be written as:

$$\sum_{j \in J_2 \cup J_1} K_j \le \min\left(\gamma L, 1 - \frac{1}{\rho} + \frac{1}{1 - \frac{1-\rho}{\sigma^2} \min_{j \in J_2} \left(\bar{P}_j / \beta_j \right)} \right), \text{ and } \sum_{j \in J_3} K_j = 0.$$

If J_2 is a null set, since the power constraints for classes in J_1 are satisfied automatically, the user capacity constraint is given accordingly by:

$$\sum_{j \in J_1} K_j \le \gamma \cdot L, \text{ and } \sum_{j \in J_3} K_j = 0.$$

4. User Capacity and Effective Bandwidth for (M,N) Decorrelator

Suppose there are N unknown users with power A_k, $k = M+1, \cdots M+N$. Let N be fixed and let total power be $\sum_{k=M+1}^{M+N} A_k^2 = \sum_{k=M+1}^{M+N} P_k = P_u$. Assume the system has control over the M known users toward satisfying their SIR requirement but does not have control over the N unknown users. The M known users are admissible into the system if there exists a power allocation scheme for the M users such that the SIR requirement of all the M known users are satisfied. The maximum admissible number of known users per degree of freedom without power constraint is called the user capacity. The effective bandwidth of a known user can be defined such that the SIR requirements of all the known users can be met if and only if the sum of the effective bandwidths of the known users is less than the total number of degrees of freedom in the system.

Recall Eq. (10), for the (M, N) decorrelator, it provides that

$$\begin{aligned} \mathrm{SIR}_i &= \frac{A_i^2}{\frac{\sigma^2}{1-\rho} \cdot \left[\frac{1+\rho(M-2)}{1+\rho(M-1)} \right] + \left[\frac{\rho}{1+\rho(M-1)} \right]^2 \sum_{k=M+1}^{M+N} A_k^2} \\ &= \frac{P_i}{\sigma^2 + (M-1) \frac{\sigma^2 \rho^2}{[1+\rho(M-1)](1-\rho)} + \left[\frac{\rho}{1+\rho(M-1)} \right]^2 \sum_{k=M+1}^{M+N} P_k}. \end{aligned} \tag{26}$$

In this equation, the three terms in the denominator are identified to be respectively (i) noise, (ii) interference from the other M-1 known users except user *i*, and (iii) interference from the N unknown users. We can identify the effective interference to be:

$$I_{ji}=\begin{cases}\dfrac{\sigma^2\rho^2 L}{[1+\rho(M-1)](1-\rho)} & j\neq i,\ j=1,2,\cdots M\\ \left[\dfrac{\rho}{1+\rho(M-1)}\right]^2 LP_j & j=M+1,\cdots,M+N\end{cases}. \tag{27}$$

Notice that the effective interference from the known users is the same as in the case of no unknown users, and the effective interference of the unknown users is proportional to their power.

Suppose there is no power constraint and all known users have the same SIR requirement β. Since all the users are symmetric, the obvious power allocation is given by

$$P_i=\beta\frac{\sigma^2}{1-\rho}\left[\frac{1+\rho(M-2)}{1+\rho(M-1)}\right]+\beta\left[\frac{\rho}{1+\rho(M-1)}\right]^2 P_u,\ i=1,2,\cdots,M\ . \tag{28}$$

Since $\beta\sigma^2<P_i<\frac{\beta\sigma^2}{1-\rho}+\beta\rho^2 P_u<\infty$, it follows that there is no constraint imposed on the number of users from the power point of view. Then $M\leq\gamma L-N$, and the user capacity is given by:

$$C_{\text{dec}}^{(M,N)}(\beta)=\gamma-\frac{N}{L}\ \text{users / degree of freedom.} \tag{29}$$

The user capacity of the (M, N) case is $\frac{N}{L}$ less than the (K, 0) case, and this is caused by the presence of the additional unknown users.

Now let us add a power constraint, namely $P_i\leq\overline{P},\ i=1,2,\cdots,M$. $(\overline{P}\geq\beta\sigma^2)$. Then we have to compare the possible values of P_i with $\overline{P}$ to obtain the limit on M. Define $\Delta(M)$ and $F(M)$ as follows

$$\Delta(M)\equiv\frac{1}{1+\rho(M-1)}, \tag{30}$$

and $F(M)\equiv-\frac{\sigma^2\rho}{1-\rho}\Delta(M)+\rho^2P_u\Delta^2(M)=\rho^2P_u\Delta(M)(\Delta(M)-2\Delta^*)$. (31)

Then we have $P_i=\beta\left[\frac{\sigma^2}{1-\rho}+F(M)\right]$. $F(\Delta)$ is a quadratic function of Δ with a minimum at $\Delta^*=\frac{\sigma^2}{2\rho(1-\rho)P_u}$. The quantity $F(M)$ has 3 possible behaviors for $\Delta^*\geq1$, $\frac{1}{2}\leq\Delta^*<1$, and $0<\Delta^*<\frac{1}{2}$, as shown in Figure 2.

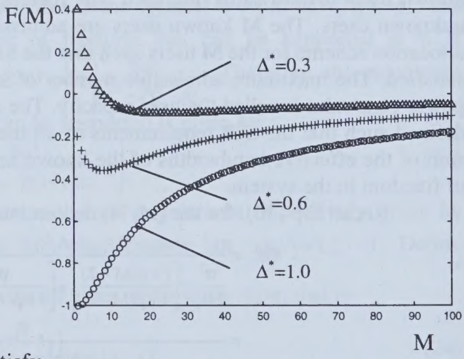


Figure 2:
Function F(M) for different values of the parameter Δ^*.
(Unit of the y-axis is ρ^2P_u)

Therefore, the user capacity has to satisfy

$F(M)\leq\frac{\overline{P}}{\beta}-\frac{\sigma^2}{1-\rho}$ and $M\leq\gamma L-N$.

The exact numerical expression of the user capacity according to the value of Δ^* was computed but is too complicated to be depicted here.

An explanation for the forms of the curves in Figure 2 is as follows. In the denominator of Eq. (26), there are 2 terms. The first term $\frac{\sigma^2[1+\rho(M-2)]}{[1+\rho(M-1)](1-\rho)}$ arises from the noise and the M-1 known other users; this term is monotonically increasing with M. The second term, $\left[\frac{\rho}{1+\rho(M-1)}\right]^2 P_u$, arises from the K-M unknown users; this term is decreasing monotonically. For a fixed number (and power) of unknown users, if we let the number of known users increase, the ratio of the power from the unknown users to the power of known users is decreasing, looks like the power of the unknown users is decreasing. Therefore, the curve shape is determined by the relative coefficients of the increasing term and the decreasing term, which is actually the parameter Δ*. When $\Delta^* \ge 1$, i.e., $\frac{\sigma^2\rho}{1-\rho} \ge \rho^2 P_u$, the increasing term is dominant, and therefore the total interference and noise is increasing with M. We want M be less than some number to satisfy the SIR requirement. When $\Delta^* < 1$ the decreasing term is dominant for small M and the increasing term becomes dominant as M becomes large. Therefore the power needed to satisfy the SIR requirement will decrease first as M increases, and then start increasing when M becomes large enough. So we may want M be less than one number to limit the increasing noise term, and (or) be larger than another number to counter balance the interference power of the unknown users.

Now suppose there are J classes, with user numbers $M_j, j=1,2,\cdots J$, ($\sum_{j=1}^{J} M_j = M$) and specific SIR requirement $SIR_j \ge \beta_j, j=1,2,\cdots J$. Define $\bar{\beta} = \max_j \beta_j$. The power allocation is the same for all users in the same class, i.e.

$$P_j = \frac{\beta_j \sigma^2}{1-\rho}\left[\frac{1+\rho(M-2)}{1+\rho(M-1)}\right] + \beta_j P_u \left[\frac{\rho}{1+\rho(M-1)}\right]^2, \quad j=1,2,\cdots J. \qquad (32)$$

If there is no power constraint, since $P_i < \frac{\bar{\beta}\sigma^2}{1-\rho} + \bar{\beta} P_u \rho^2 < \infty$, there is no constraint on the number of users from the power point of view, and the user capacity constraint is given by $\sum_{j=1}^{J} M_j \le \gamma L - N$. Thus, we can identify the effective bandwidth as

$$e_{dec}^{(M,N)}(\beta) = \frac{1}{\gamma - \frac{N}{L}} \quad \text{degree of freedom / user}. \qquad (33)$$

This effective bandwidth clearly is larger than $\frac{1}{\gamma}$, which is the effective bandwidth for the (K, 0) decorrelator.

With a power constraint, the user capacity constraint is complicated; it depends on Δ*. When $\Delta^* \ge 1$, similarly to the case of the (K, 0) decorrelator; we need to partition the J classes into 3 sets. When $0 < \Delta^* < \frac{1}{2}$, or $\frac{1}{2} \le \Delta^* < 1$, it is even more complicated, and we need to partition the J classes into 4 sets. We omit the detailed discussion of this case here.

5. Discussion and Conclusions

There are some restrictions on the values of L, K and ρ. They can not be arbitrary.

Restriction 1: In L-dimensional space, K ($K \le L$) symmetric vectors with unit length and crosscorrelation ρ exist, if and only if $-\frac{1}{K-1} \le \rho \le 1$.

Restriction 2: In L-dimensional space (L finite), the maximum number of symmetric vectors is L+1, and then the value of ρ must be equal to $-\frac{1}{L}$. In another word, $K \le L+1$, with equality only when $\rho = -\frac{1}{L}$.

We omit the proof here. From these two restrictions, if K=L+1, the only possible value of ρ is $\rho = -\frac{1}{L}$; if $K \le L$, the possible values of ρ satisfy $-\frac{1}{K-1} \le \rho \le 1$. If we keep $\frac{K}{L} = \alpha \le 1$, notice that the values $\frac{1}{L}$, or $\frac{1}{\sqrt{L}}$ that we used for ρ in section 3 are inside the region $[-\frac{1}{K-1},1]$. Negative values of ρ have very limited range, namely the interval $[-\frac{1}{K-1},0]$, which vanishes as K gets large. So we consider that $0 \le \rho \le 1$ in our calculations. In section 4, we assume that " in an L-dimensional space, there are at most γL symmetric users with crosscorrelation ρ." From the above explanation, for finite L, the value of γ should be 1.

In conclusion, for the (K, 0) decorrelator, the user capacity is 1, and the effective bandwidth is 1. For the (M, N) decorrelator with fixed unknown power, the user capacity is $1-\frac{N}{L}$ and the effective bandwidth is $\frac{1}{1-\frac{N}{L}}$, which goes to 1 as L becomes large.

There are many interesting and realistic subsequent related questions that arise from the approach we have taken, such as other multiuser detector and using random sequences. This preliminary investigation will help address them more effectively in the future.

References:

[1] Sergio Verdú, *Multiuser Detection*. Cambridge Univ. Press, 1998.

[2] David N. C. Tse and Stephen V. Hanly, "Linear multiuser receivers: effective interference, effective bandwidth and user capacity," *IEEE Trans. Inform. Theory*, vol. 45, No. 2, pp. 641-657, March 1999.

[3] Pramod Viswanath, Venkat Anantharam, and David N. C. Tse, "Optimal sequence, power control, and user capacity of synchronous CDMA systems with linear MMSE multiuser receivers," *IEEE Trans. Inform. Theory*, vol. 45, No. 6, pp. 1968-1983, Sept. 1999.

[4] Roch Guérin, Hamid Ahmadi, and Mahmoud Naghshineh, "Equivalent capacity and its application to bandwidth allocation in high-speed networks," *IEEE J. Select. Areas Commun.*, vol. 9, No. 7, pp.968-981, September 1991.

[5] S. Verdú and S. Shamai (Shitz), "Spectral efficiency of CDMA with random spreading", *IEEE Trans. Inform. Theory*, vol. 45, pp. 622-640, Mar. 1999.

[6] Junshan Zhang and Edwin K. P. Chong, " CDMA systems in fading channels: admissibility, network capacity, and power control," *IEEE Trans. Inform. Theory*, vol. 46, No. 3, pp. 962-981, May 2000.

[7] S. Verdú, " Optimum multiuser asymptotic efficiency", *IEEE Trans. Commun.*, vol. 34, pp. 890-897, Sept. 1996.

[8] R. Lupas and S. Verdú, "Linear multiuser detectors for synchronous code-division multiple access," *IEEE Trans. Inform. Theory*, vol. 35, pp. 126-136, Jan. 1989.

[9] R. Lupas and S. Verdú, "Near-far resistance of multiuser detectors in asynchronous channels," *IEEE Trans. Commun.*, vol. 38, pp. 496-508, Apr. 1990.

[10] Z. Xie, Short, and C. Rushforth, " A family of suboptimum detectors for coherent multiuser communications," *IEEE J. Select. Areas Commun.*, vol. 8, pp.683-690, May 1990.

[11] U. Madhow and M. Honig, "MMSE interference suppression for direct sequence spread-spectrum CDMA," *IEEE Trans. Commun.*, vol. 42, pp. 3178-3188, Dec. 1994.

[12] U. Madhow and M. Honig, "On the average near-far resistance for MMSE detection of direct sequence CDMA signals with random spreading," *IEEE Trans. Inform. Theory*, vol. 45, pp. 2039-2045, Sept. 1999.

A Novel Suboptimal CDMA Multiuser Detector

H. F. Francis Lu
Dept. of Electrical Engineering
University of Southern California
Los Angeles, CA 90089
hsiaofel@scf.usc.edu

P. Vijay Kumar
Dept. of Electrical Engineering
University of Southern California
Los Angeles, CA 90089
vijayk@ceng.usc.edu

Abstract

In this paper we propose a novel 2-stage multiuser receiver. The first stage invokes Choleskey factorization of the correlation matrix and a form of decision feedback to provide an estimate of data sequences. The second stage attempts to improve performance by perturbing this estimate and choosing the perturbation closest to the received signal. The proposed receiver has much less complexity than the optimal detector. Simulation results are presented to show that the proposed detector has almost the same BER performance as the optimal Verdú receiver.

1 Introduction

Suppose there are K users in the system and each user transmits M bits of information via a quasi-static additive white Gaussian noise channel. We further assume the following are known to the receiver: the uniform bit energy E_b, fading amplitude A_k of k^{th} user introduced by channel, and time delay T_k of k^{th} user, for all K users. Therefore, the received waveform is $r(t) = \sum_{k=1}^{K} \sum_{m=1}^{M} y_k(m) A_k S_k(t - mT - T_k) + w(t)$ where $y_k(m) \in \mathcal{P}$ is the m^{th} transmitting bit of k^{th} user; $\mathcal{P} = \left\{+\sqrt{E_b}, -\sqrt{E_b}\right\}$; $S_k(t)$ is the signature waveform of k^{th} user with time duration T. $w(t)$ is the additive white Gaussian noise with double-sided power spectral density $\frac{N_0}{2}$.

Define $\underline{b}^T \triangleq [b_1(1), b_2(1), \ldots, b_K(1), b_1(2), \ldots b_K(M)]$, then the squared Euclidean distance $\mathcal{D}_{\underline{b}}$ between received waveform $r(t)$ and the bit-stream waveform of any sequence $\underline{b}$ is given by $\mathcal{D}_{\underline{b}} \triangleq \int_{-\infty}^{\infty} \left| r(t) - \sum_{k=1}^{K} \sum_{m=1}^{M} A_k b_k(m) S_k(t - mT - T_k) \right|^2 dt$. The maximum-likelihood receiver attempts to find $\underline{b}$ such that $\mathcal{D}_{\underline{b}}$ is the minimum.

Define $\underline{r}^T \triangleq [r_1(1), r_2(1), \ldots, r_K(M)]$ where $r_k(m) \triangleq \int_{-\infty}^{\infty} r(t) A_k S_k(t - mT - T_k)\, dt$. To simplify notation, we make the following identification $A_i \triangleq A_k$, $b_i \triangleq b_k(m)$, $r_i \triangleq r_k(m)$, $S_i(t) \triangleq S_k(t - mT - T_k)$, where $i = m \cdot K + k$. Define the positive-definite symmetric correlation $(KM \times KM)$ matrix $\mathbf{H}$ by

$H_{ij} \triangleq \int_{-\infty}^{\infty} A_i A_j S_i(t) S_j(t)\, dt$. Then $\hat{\underline{b}} = \arg\min_{\underline{b}} \mathcal{D}_{\underline{b}} = \arg\min_{\underline{b}} \left\{-2\underline{b}^T \underline{r} + \underline{b}^T \mathbf{H} \underline{b}\right\}$ and $\underline{r}$ is a sufficient statistic for optimal multiuser detection[1][2]. It is known that the (optimal) multiuser detector that solves this minimization problem has exponential complexity [2].

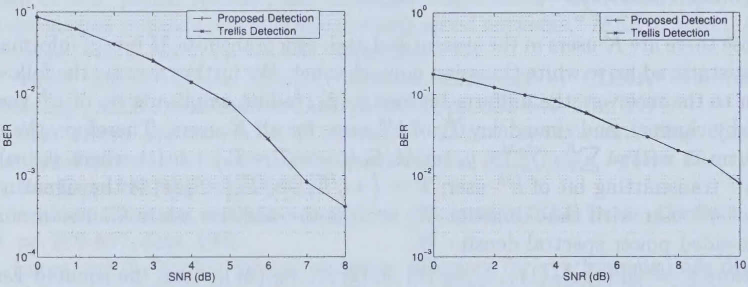
2 Detection based on Choleskey Factorization

Since $\mathbf{H}$ is a positive definite symmetric matrix, using Choleskey factorization [3] we have $\mathbf{H} = \mathbf{L}\mathbf{L}^T$ where $\mathbf{L}$ is a lower triangular matrix. Let $\underline{r} = \mathbf{H}\underline{x}$ and we can then rewrite $\arg\min_{\underline{b}} \left\{-2\underline{b}^T\underline{r} + \underline{b}^T\mathbf{H}\underline{b}\right\} = \arg\min_{\underline{b}} \left\|\mathbf{L}^T\left(\underline{x} - \underline{b}\right)\right\|^2$. For simplicity, let $\mathbf{F} \triangleq \mathbf{L}^T$. Since $\left\|\mathbf{F}\left(\underline{x} - \underline{b}\right)\right\|^2 = \sum_{q=1}^{KM} \left|\mathbf{F}_q\left(\underline{x} - \underline{b}\right)\right|^2$, where $\mathbf{F}_q$ is the q^{th} row of $\mathbf{F}$, in the first stage we detect bit b_q in the following way such that the q^{th} term is minimum. This is possible since the matrix $\mathbf{F}$ is an upper-triangular matrix and b_q has no contribution to the $(q+1)^{th}$ term.

First Stage: Set $\hat{b}_{KM}^1 = \sqrt{E_b}$, $\hat{b}_{KM}^2 = -\sqrt{E_b}$. Assuming the $q+1, q+2, \ldots, KM$ bits had been detected, the decoding rule of the q^{th} bit is $\hat{b}_q^i = \arg\min_{\hat{b_q} \in \mathcal{P}} \left\|\mathbf{F}_q\left(\underline{x} - \underline{\hat{b}_j^i}\right)\right\|$, for each i. Repeat previous step until all data is recovered. $\underline{\hat{b}} = \arg\min_{\underline{b} \in \left\{\underline{\hat{b}^1}, \underline{\hat{b}^2}\right\}} \left\|\mathbf{F}\left(\underline{x} - \underline{b}\right)\right\|$.

Second Stage: Update $\hat{b}_{KM}$ using $\hat{b}_{KM} = \arg\min_{\hat{b}_{KM} \in \mathcal{P}} \left(\underline{x} - \underline{\hat{b}}\right)^T \mathbf{H} \left(\underline{x} - \underline{\hat{b}}\right)$. Having updated $\hat{b}_{q+1}, \ldots,$ and $\hat{b}_{KM}$, update $\hat{b}_q$ using $\hat{b}_q = \arg\min_{\hat{b}_q \in \mathcal{P}} \left(\underline{x} - \underline{\hat{b}}\right)^T \mathbf{H} \left(\underline{x} - \underline{\hat{b}}\right)$. Repeat previous step until $\hat{b}_1$ has been updated. Repeat whole process until iteration converges.

In order to have better performance, using exhaustive search over every t consecutive bits is also possible. But as can be seen from the simulation, even the setting $t = 1$ is sufficient to yield a detector with error performance close enough to the optimal detector. Setting $t = 2$ does not significantly improve performance.



(a) (b)

Figure 2: K=6 M=20 using Gold sequences with period 31 over (a) nonfading channel and (b) fading channel

References

[1] S. Verdú, "Optimum Multi-user Asymptotic Efficiency," *IEEE Trans. on Communications*, vol. COM-34, no. 9, pp. 980-897, Sept. 1986

[2] S. Verdú, *Multiuser Detection*, Cambridge University Press, 1998.

[3] C. Schlegel and L. Wei, "A Simple Way to Compute the Minimum Distance in Multiuser CDMA Systems," *IEEE Trans. on Communications*, vol. COM-45, no. 5, pp. 532-535, May 1997.

Blind Adaptive Linear Multiuser Detection for Variable Spreading Factor CDMA

Neal Seidl
EECE Department
Marquette University
neal.seidl@marquette.edu

Ivan Howitt
EECS Department
University of Wisconsin, Milwaukee
howitt@ee.uwm.edu

James Richie
EECE Department
Marquette University
richiej@marquette.edu

1. Introduction

In variable-spreading-factor (VSF) multirate CDMA systems, users transmit using the same chip rate and employ different processing gains in order to achieve different data rates. The linear minimum mean-square error (LMMSE) detector for VSF-CDMA was shown to be a periodically time varying (PTV) filter in [1-4]. All of these previously proposed multirate detectors are based on a translation series representation of the PTV filter. In this paper we propose a new detector based on a harmonic series representation (HSR). The new detector operates adaptively without training data requiring only the timing and spreading code of the user of interest.

2. System Model

We consider the baseband asynchronous VSF direct-sequence CDMA system where K users simultaneously transmit binary data over an AWGN channel. The received signal is given by

$$r(t)=\sum_{k=1}^{K} A_k \sum_{i=-\infty}^{\infty} b_k^{(i)} s_k(t-iT_k-\tau_k)+n(t) \tag{1}$$

where $b_k^{(i)} \in \{\pm 1\}$ is the kth user's ith data bit assumed to be independent and equiprobable, A_k is the kth user's received amplitude, T_k is the kth user's bit duration, $\tau_k \in [0,T_k]$ is the kth user's delay, and $n(t)$ is AWGN. The signature waveform for user k is

$$s_k(t)=\sum_{n=0}^{N_k-1} s_k^{(n)}\Psi(t-nT_c) \tag{2}$$

where T_c is the chip duration, $N_k = T_k / T_c$ is the kth user's processing gain assumed to be an integer-power of 2, $s_k^{(n)} \in \{\pm 1\}$ is the nth chip in the kth user's spreading code, and $\Psi(t)$ is the unit-energy chip waveform with support confined to $[0,T_c]$.

The second order statistics of $r(t)$ are periodic with period pT_1, where p is the smallest integer such that pT_1/T_k is a positive integer for all $1<k\le K$. Since all processing gains $\{N_k\}$ are integers and all users have the same chip rate, all ratios $\{T_1/T_k\}$ are rational and the period pT_1 will be finite.

3. Proposed Detector

The proposed detector processes a discrete-time version of the received signal obtained by chip-matched filtering and subsequent chip-rate sampling synchronized to the user of interest. The received signal is processed in finite windows of length L centered about the bit of interest. The HSR detector structure is depicted in Figure 1. For the ith bit it decides $\hat{b}^{(i)} = \mathrm{sgn}\left[\Re\left(y^{(i)}\right)\right]$ where the decision statistic $y^{(i)}$ is given by

$$y^{(i)}=\sum_{n=0}^{p-1} \mathbf{h}_n^H \mathbf{r}^{(i)} e^{j2\pi in/p} \tag{3}$$

where the complex L-vector $\mathbf{h}_n$ is the impulse response of the nth "branch" of the detector, and $\mathbf{r}^{(i)}$ is the vector of samples from the processing window for the ith bit.

The p detector branches are chosen to minimize the energy in the decision statistic with $\mathbf{h}_0$ constrained so that desired signal can not be canceled. That is,

$$\min_{\{\mathbf{h}_n\}}\left(E\left[\left|y^{(i)}\right|^2\right]\right), \quad \text{subject to}: \mathbf{h}_0^T\bar{\mathbf{s}}_1 = 1 \tag{4}$$

where $\bar{\mathbf{s}}_1$ is the user of interest's spreading code symmetrically zero-augmented to length L. Using stochastic gradient descent, the branches may be obtained adaptively as follows,

$$\mathbf{h}_n^{(i+1)} = \begin{cases} \mathbf{h}_n^{(i)} - \mu\, y^{(i)}\left(\mathbf{r}^{(i)} - \bar{\mathbf{s}}_1^T\mathbf{r}^{(i)}\bar{\mathbf{s}}_1\right), & n = 0 \\ \mathbf{h}_n^{(i)} - \mu\, y^{(i)}\mathbf{r}^{(i)}e^{j2\pi in/p}, & n \geq 1 \end{cases}, \tag{5}$$

where $\mathbf{h}_n^{(i)}$ is the estimate of $\mathbf{h}_n$ at iteration i, and μ is the LMS step size. The same *constrained minimum output energy* (CMOE) adaptation applied to the time invariant detector was shown in [5] to converge to the LMMSE detector for single-rate CDMA.

4. Simulation Example

Monte Carlo simulation was performed comparing the new HSR multirate detector to the single-rate adaptive detector from [5]. The user of interest employs a pocessing gain of $N_1 = 8$. There are three ½-rate interferers persent at the beginning, and at iteration 1000 a new ¼ interferer is added. All Interferers are at +10dB, and the SNR was 10 dB. The delays were independent and uniformly distributed, and the results shown are ensemble averages from 100 runs. The HSR detector used $p = 4$ branches. The results are shown in Figure 2. The improvement over the time invariant detector under multirate interference is obvious.

References

[1] H. Ge and J. Ma, "Multi-Rate LMMSE Detectors for Asynchronous Multi-Rate CDMA Systems," presented at IEEE International Conference on Communications, 1998.

[2] A. Buzzi, M. Lops, and A. Tulino, "Linear-Conjugate/Linear Filtering for Interference Rejection in Multi-Rate DS/CDMA Systems," presented at IEEE WCNC, New Orleans, LA, 1999.

[3] A. Sabharwal, U. Mitra, and R. Moses, "Cyclic Wiener filtering based multirate DS-CDMA Receivers," presented at IEEE WCNC, New Orleans, LA, 1999.

[4] N. Seidl, I. Howitt, and J. Richie, "Blind Adaptive Linear Multiuser Detection for Multirate CDMA Systems," presented at Vehicular Technology Conference, Boston, MA, 2000.

[5] M. Honig, U. Madhow, and S. Verdú, "Blind Adaptive Multiuser Detection," *IEEE Transactions on Information Theory*, vol. 41, pp. 944-960, 1995.

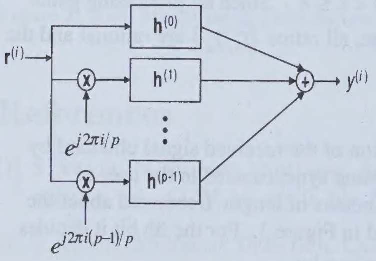


Figure 1: HSR detector structure.

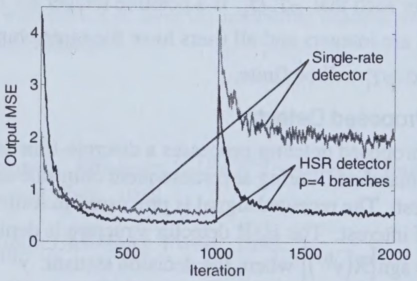


Figure 2: Adaptive Detector Convergence

On the Number of Samples Required for Q-Learning*

V. S. Borkar

School of Technology and Computer Science
Tata Institute of Fundamental Research
Homi Bhabha Road, Mumbai 400005, INDIA
borkar@tifr.res.in

Abstract

For the synchronous Q-learning algorithm for discounted cost control of finite Markov chains, an estimate of the additional number of iterates required to be within prescribed accuracy of the optimum with high probability is obtained, using the o.d.e. approximation and concentration inequalities.

1 Introduction

In a recent work [3], the author analysed the 'lock-in' probability of a stochastic approximation algorithm, i.e., the probability of its converging to a given equilibrium if it is in its domain of attraction after sufficiently many iterates. The present work specializes this argument to the Q-learning algorithm to get a result in the spirit of sample complexity. To be precise, given a large and a small neighbourhood of the desired limit and the fact that the algorithm is found inside the former after sufficiently many iterates, we prescribe the number of additional iterates required to hit the latter neighbourhood (and remain there) with a prescribed probability. This is not a true sample complexity result because of the qualification 'sufficiently large' above, but it does give some useful information.

The next section introduces the Q-learning algorithm [11] for discounted cost control of finite Markov chains. Section 3 contains the main results.

2 The Q-learning algorithm

Consider a controlled Markov chain $X_n, n \geq 0$, on a finite state space S, with a finite action space A, and transition probabilities $p(i, a, j) \in [0, 1],\ i, j \in S,\ a \in A$, satisfying $\sum_j p(i, a, j) = 1$. Let $\{Z_n\}$ denote the A-valued control process. Thus for $n \geq 0$,

$$P(X_{n+1} = j / X_m, Z_m, m \leq n) = p(X_n, Z_n, j),\ j \in S.$$

Let $\beta \in (0, 1), k : S \times A \to R$ be resp. the discount factor and the running cost function, the objective being to minimize over all admissible controls the total discounted cost

---

*This research is supported in part by grant no. III 5(12)/96-ET from the Department of Science and Technology, Government of India.

$E[\sum_{m=0}^{\infty} \beta^m k(X_m, Z_m)]$. Let $K > 0$ be a bound on $|k|$ and $d = |S| \times |A|$. Define the value function $V : S \to R$ by

$$V(i) = min E[\sum_n \beta^n k(X_n, Z_n)/X_0 = i],$$

where the minimization is over all admissible controls. Then V is the unique solution to the dynamic programming equations [8]

$$V(i) = min_a[k(i,a) + \beta \sum_j p(i,a,j)V(j)], i \in S. \tag{1}$$

Let $Q(i,a)$ denote the expression in the square brackets in (1). Then $V(i) = min_a Q(i,a)$ and $Q = [Q(i,a)]$ is the unique solution in R^d to

$$Q(i,a) = k(i,a) + \beta \sum_j p(i,a,j) min_b Q(j,b), \ i \in S, \ a \in A. \tag{2}$$

In fact, define $F = [F_{ia}]_{i \in S, a \in A} : R^d \to R^d$ by :

$$F_{ia}(x) = k(i,a) + \beta \sum_j p(i,a,j) min_b x_{jb}, \ i \in S, \ a \in A,$$

for $x = [x_{ia}] \in R^d$. Then (2) is of the form $Q = F(Q)$. Letting $||x||_\infty = max_{i,a}|x|$ for $x = [x_{ia}] \in R^d$, it is easily verified that $||F(x) - F(y)||_\infty \leq \beta ||x - y||_\infty, \ x, y \in R^d$. Thus F is an $|| \cdot ||_\infty$-contraction and hence has a unique fixed point Q by the contraction mapping principle. Furthermore, $Q^n, n \geq 0$, given iteratively by

$$Q^{n+1}(i,a) = k(i,a) + \beta \sum_j p(i,a,j) min_b Q^n(j,b), \ i \in S, a \in A, \tag{3}$$

with arbitrary Q^0, will converge exponentially to Q.

Suppose now that $p(\cdot,\cdot,\cdot)$ is unknown, but one has at hand a simulation device which can simulate an S-valued random variable ξ with law $p(i,a,\cdot)$ for any given i, a. The Q-learning algorithm then replaces the conditional average w.r.t. $p(i,a,\cdot)$ by an evaluation at an independently simulated ξ as above. One then incorporates incremental moves as in stochastic approximation so that the algorithm 'sees' conditional average in question via the well-known averaging properties of stochastic approximation. To be precise, the algorithm is :

$$Q^{n+1}(i,a) = Q^n(i,a) + a(n)(k(i,a) + \beta min_b Q^n(\xi_{ia}^n, b) - Q^n(i,a)), \ i \in S, a \in A, \tag{4}$$

where $a(n) > 0$ satisfy :

$$\sum_n a(n) = \infty, \ \sum_n a(n)^2 < \infty, \tag{5}$$

and ξ_{ia}^n is a random variable with law $p(i,a,\cdot)$ generated by the simulation device independent of all other random variables generated up to time n.

This is the *synchronous* version of the algorithm because it updates all components of Q^n simultaneously at each iteration, using the most recent values available thereof. Of much greater interest is the *asynchronous* version wherein only some components get updated at each time, each by a different processor with its own 'clock', and these processors communicate with each other with communication delays. In this paper we stay with the synchronous case for sake of simplicity and remark briefly on the asynchronous case at the end.

The Q-learning algorithm has been extensively analysed in [1], [4], [6], [9], [10], [12]. In particular the a.s. boundedness of the iterates and their a.s. convergence to the desired limit is established. One approach [4] is the o.d.e. (for *ordinary differential equations*) approach which treats (4) as a noisy discretization of the o.d.e.

$$\dot{q}(t) = h(q(t)), \tag{6}$$

where $h(q) = F(q) - q, q \in R^d$. Note that

$$||h(x) - h(y)||_\infty \leq 2||x - y||_\infty, \; x, y \in R^d, \tag{7}$$

implying in particular that (6) is well-posed. In fact, (6) is a special case of a much more general class of o.d.e.s studied in [5] where it is shown that

$$||q(t) - Q||_\infty \leq e^{-\beta(t-s)}||q(s) - Q||_\infty, \; t > s. \tag{8}$$

We shall use this fact later. For notational ease, write $|q|^* = ||q - Q||_\infty$.

3 Main Results

Let $C >> 2\epsilon > 0$ and suppose that $|Q^{n_0}|^* < C$ for n_0 'sufficiently large'. (The latter qualification is made precise later.) Our aim will be to estimate the number N of additional iterates required to ensure that $|Q^n|^* \leq 2\epsilon$ for $n \geq n_0 + N$ with a 'high' probability. We proceed through a sequence of lemmas. We consider the iterates for $n \geq n_0$ conditioned on Q^{n_0} and thus may view the latter as deterministic by abuse of notation. Let $T > 0, a = sup_{n \geq n_0} a(n)$, $\bar{T} = T + a, T_0 = t(0) = 0$, $t(n) = \sum_{m=n_0}^{n_0+n} a(m)$, $T_n = min\{t(m) : t(m) \geq T_{n-1} + T\}$. Then $T_{n+1} \in [T_n + T, T_n + \bar{T}]$ and $T_n = t(m(n))$ for suitable $m(n) \uparrow \infty$. Also, $t(n), T_n \uparrow \infty$ as $n \uparrow \infty$. Define $\bar{q}(t), t \geq 0$, by : $\bar{q}(t(n)) = Q^{n_0+n}$, with linear interpolation on each $[t(n), t(n+1)]$. Also define $\tilde{q}(t), t \geq 0$, by $\tilde{q}(t) = Q^{n_0+n}, t \in [t(n), t(n+1)), n \geq 0$. Finally, let $q^n(t), t \in I_n \triangleq [T_n, T_{n+1}]$, be the solution to (6) with $q^n(T_n) = \bar{q}(T_n)$.

Lemma 3.1 For $t \in I_n$, $|\bar{q}(t)|^* \leq K\bar{T} + |\bar{q}(T_n)|^*$.

Proof From (3), it follows that for $m(n) \leq k < m(n+1)$,

$$\begin{aligned} |\bar{q}(t(k+1))|^* &\leq (1 - a(k))|\bar{q}(t(k))|^* + a(k)(K + \beta|\bar{q}(t(k))|^*) \\ &= (1 - (1 - \beta)a(k))|\bar{q}(t(k))|^* + a(k)K. \end{aligned}$$

Iterating,

$$|\bar{q}(t(k+1))|^* \leq K\bar{T} + e^{-(1-\beta)(t(k+1)-T_n)}|\bar{q}(T_n)|^*.$$

The claim follows. □

Note that by (8), $|q^n(T_{n+1})|^* \leq e^{-\beta T}|q^n(T_n)|^*$. Let $\delta < \frac{(1-e^{-\beta T})\epsilon}{4}$, $b(n) = \sum_{m \geq n} a(n)^2$. (Thus $b(n) \downarrow 0$.)

Lemma 3.2 On $B_n \triangleq \{sup_{t \in I_m}||q^m(t) - \bar{q}(t)||_\infty \leq \delta, n_0 \leq m < n\}$, $|\bar{q}(t)|^* \leq C$ for $t \in I_{n-1}$. In particular, $|\bar{q}(T_n)|^* \leq C$ and $|\bar{q}(t)|^* \leq K\bar{T} + C$ for $t \in I_n$.

Proof Let $m < n$. If $|q^m(T_m)|^* = |\bar{q}(T_m)|^* \leq C$, then $|q^m(T_{m+1})|^* \leq e^{-\beta T}C$ and hence $|\bar{q}(T_{m+1})|^* \leq e^{-\beta T}C + \delta \leq C$. Iterating for $n_0 \leq m < n$, the first claim follows. The second claim then follows from the preceding lemma. □

Define the R^d-valued process $\zeta^n = [\zeta^n_{ia}], i \in S, a \in A, n \geq n_0$, by:

$$\zeta^n_{ia} = \sum_{m=n_0}^{n} a(m)(min_b(Q^m(\xi^m_{ia}, b) - \sum_j p(i,a,j) min_b Q^m(j,b)).$$

Then $\{\zeta^n\}$ is a d-dimensional martingale w.r.t. $\sigma(\xi^m_{ia}, i \in S, a \in A, m \leq n)$, $n \geq 0$. Rewrite (3) as : For $m(n) \leq k < m(n+1)$,

$$\begin{aligned}\bar{q}(t(k+1)) &= \bar{q}(t(k)) + \int_{t(k)}^{t(k+1)} h(\tilde{q}(t))dt + \zeta^{k+1} - \zeta^{m(n)} \\ &= \bar{q}(T_n) + \int_{T_n}^{t(k+1)} h(\bar{q}(t))dt + \int_{T_n}^{t(k+1)} (h(\tilde{q}(t)) - h(\bar{q}(t)))dt + (\zeta^{k+1} - \zeta^{m(n)}).\end{aligned}$$

Also,

$$q^n(t(k+1)) = \bar{q}(T_n) + \int_{T_n}^{t(k+1)} h(q^n(t))dt.$$

Subtracting and taking norms, we have, in view of (7),

$$\begin{aligned}||\bar{q}(t(k+1)) - q^n(t(k+1))||_\infty &\leq 2\int_{T_n}^{t(k+1)} ||\bar{q}(t) - q^n(t)||_\infty dt \\ &+ 2\int_{T_n}^{t(k+1)} ||\bar{q}(t) - \tilde{q}(t)||_\infty dt \\ &+ ||\zeta^{k+1} - \zeta^{m(n)}||_\infty.\end{aligned}$$

On B_n, $|\bar{q}(t)|^* \leq K\bar{T} + C$ for $t \in I_n$. Thus

$$||\bar{q}(t) - \tilde{q}(t)||_\infty \leq (K + 2(K\bar{T} + C))a(k),\ t \in I_k,\ m(n) \leq k < n.$$

Hence

$$\int_{t(k)}^{t(k+1)} ||\bar{q}(t) - \tilde{q}(t)||_\infty dt \leq (K + 2(K\bar{T} + C))a(k)^2.$$

Thus

$$\begin{aligned}||\bar{q}(t(k+1)) - q^n(t(k+1))||_\infty &\leq 2\int_{T_n}^{t(k+1)} ||\bar{q}(t) - q^n(t)||_\infty dt \\ &+ 2(K + 2(K\bar{T} + C))(\sum_{i=m(n)}^{m(n+1)} a(i)^2) \\ &+ sup_{m(n) \leq k < m(n+1)} ||\zeta^k - \zeta^{m(n)}||_\infty.\end{aligned}$$

By the Gronwall lemma, we then have :
Lemma 3.3 On B_n,

$$sup_{t \in I_n} ||\bar{q}(t) - q^n(t)||_\infty \leq e^{2\bar{T}}(2(K + 2K\bar{T} + 2C)b(m(n)) + sup_{m(n) \leq k < m(n+1)} ||\zeta^k - \zeta^{m(n)}||_\infty).$$

Now we specify our condition on n_0 : Let n_0 be such that

$$b(m(n_0)) \leq min(\frac{e^{-2\bar{T}}\delta}{4(K + 2K\bar{T} + 2C)}, \frac{e^{-4\bar{T}}\delta^2}{32(C + K\bar{T})^2}). \tag{9}$$

Lemma 3.4 Under (9), for $M = \frac{e^{-4\bar{T}}\delta^2}{32(C+K\bar{T})^2}$,

$$P(sup_{t \in I_n} ||\bar{q}(t) - q^n(t)||_\infty > \delta\ for\ some\ n \geq n_0) \leq 2de^{-\frac{M}{b(m(n_0))}}.$$

Proof Note that the l. h. s. equals

$$\sum_{n\geq n_0} P(sup_{t\in I_n}||\bar{q}(t) - q^n(t)||_\infty > \delta/B_n).$$

But by Lemma 3.3 and (9),

$$P(sup_{t\in I_n}||\bar{q}(t) - q^n(t)||_\infty > \delta/B_n) \leq P(sup_{m(n)\leq k<m(n+1)}||\zeta^k - \zeta^{m(n)}||_\infty > \frac{e^{-2\bar{T}}\delta}{2}/B_n).$$

By Lemma 3.2, on B_n, for $m(n) \leq k < m(n+1)$,

$$|\zeta_{ia}^k - \zeta_{ia}^{k-1}| \leq 2a(k)|Q^k|^* \leq 2a(k)(C + K\bar{T}).$$

Then by (3.30), p. 227, [7],

$$P(sup_{m(n)\leq k<m(n+1)}||\zeta^k - \zeta^{m(n)}||_\infty > \frac{e^{-2\bar{T}}\delta}{2}/B_n) \leq 2de^{-\frac{M}{b(m(n))-b(m(n+1))}}.$$

Hence

$$P(sup_{t\in I_n}||\bar{q}(t) - q^n(t)||_\infty > \delta \; for \; some \; n \geq n_0) \leq 2d\sum_{n\geq n_0} e^{-\frac{M}{b(m(n))-b(m(n+1))}}.$$

By (9), $b(m(n)) < M$. Since the function $x \to \frac{e^{-\frac{M}{x}}}{x}$ is increasing on $(0, M)$, the above sum is bounded by

$$2d\sum_{n\geq n_0}(b(m(n)) - b(m(n+1)))\frac{e^{-\frac{M}{b(m(n))-b(m(n+1)))}}}{b(m(n)) - b(m(n+1))} \leq 2de^{-\frac{M}{b(m(n_0))}}.$$

Our main result is :
Theorem 3.1 For n_0 satisfying (9), with probability exceeding $1 - 2de^{-\frac{M}{b(m(n_0))}}$, one has $|Q^n|^* \leq 2\epsilon$ for $n \geq n_0 + N$, where

$$N = min\{n : \sum_{i=n_0}^{n} a(i) \geq \frac{2C\bar{T}}{(1-e^{-\beta T})\epsilon}\}.$$

Proof By the foregoing, with probability exceeding $1 - 2de^{-\frac{M}{b(m(n_0))}}$, $sup_{t\in I_n}||\bar{q}(t) - q^n(t)||_\infty \leq \delta$ for $n \geq n_0$. If $|\bar{q}(T_n)|^* \geq \epsilon$, then

$$|q^n(T_{n+1})|^* \leq |q^n(T_n)|^* e^{-\beta T} \leq |q^n(T_n)| - (1 - e^{-\beta T})\epsilon.$$

Therefore

$$\begin{aligned} |\bar{q}(T_{n+1})|^* &\leq |\bar{q}(T_n)|^* - (1 - e^{-\beta\bar{T}})\epsilon + \delta \\ &\leq |\bar{q}(T_n)|^* - \frac{1}{2}(1 - e^{-\beta\bar{T}})\epsilon. \end{aligned}$$

Thus in at most $N^* = \frac{2C}{(1-e^{-\beta\bar{T}})\epsilon}$ steps, $|\bar{q}(T_n)|^* \leq \epsilon$. If $|\bar{q}(T_n)|^* \leq \epsilon$, then either $|\bar{q}(T_{n+1})|^* \leq \epsilon$ or $\epsilon < |\bar{q}(T_{n+1})|^* \leq \epsilon + \delta$. In the latter case, the above argument shows that $|\bar{q}(T_{n+2})|^* \leq \epsilon$. Since $2\delta < \epsilon$, combining the facts that $q^n(T_n) = \bar{q}(T_n)$, $q^n(t) \downarrow$ on I_n and $||q^n(t) - \bar{q}(t)||_\infty < \delta$ on I_n, it follows that with probability exceeding $1 - 2de^{-\frac{M}{b(m(n_0))}}$, $|Q^n|^* \leq 2\epsilon$ for $n \geq n_0 + N$, where $N = min\{n : \sum_{i=n_0}^{n} a(i) \geq N^*\bar{T}\}$. □

The results can be extended to the asynchronous case under reasonable assumptions. One has to use the appropriate extension of the o.d.e. approach to the asynchronous case, e.g., as in [2]. The bound one gets is similar except for an additional scalar factor.

References

[1] J. Abounady, V. S. Borkar and D. Bertsekas, "Stochastic approximation for nonexpansive maps : applications to Q-learning algorithms", *submitted*, 1998.

[2] V. S. Borkar, "Asynchronous stochastic approximation", *SIAM Journal of Control and Optimization*, vol. 36, no. 3, pp. 840-851, May 1998. (Correction note in *ibid.*, vol. 38, no. 2, pp. 662-663, Jan.-Feb. 2000.)

[3] V. S. Borkar, "On the lock-in probability of stochastic approximation", *submitted*, May 2000.

[4] V. S. Borkar and S. P. Meyn, "The O.D.E. method for convergence of stochastic approximation and reinforcement learning", *SIAM Journal of Control and Optimization*, vol. 38, no. 2, pp. 447-469, Jan.-Feb. 2000.

[5] V. S. Borkar and K. Soumyanath, "An analog scheme for fixed point computation, part I : Theory", *IEEE Trans. on Circuits and Systems - I: Fundamental Theory and Applications*, vol. 44, no. 4, pp. 351-355, April 1997.

[6] T. Jaakola, M. Jordan and S. P. Singh, "On the convergence of stochastic iterative dynamic programming algorithms", *Neural Computation*, vol. 6, no. 6, pp. 1185-1201, 1994.

[7] C. McDiarmid, "Concentration", in *Probabilistic Methods for Algorithmic Discrete Mathematics*, M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, B. Reed (eds.), pp. 195-248, Berlin-Heidelberg: Springer Verlag, 1998.

[8] M. Puterman, "Markov Decision Processes", New York: John Wiley, 1994.

[9] C. Szepesvári and M. Littman, "A unified analysis of value-function-based reinforcement-learning algorithms", *Neural Computation*, vol. 11, no. 8, pp. 2017-2060, Nov. 1999.

[10] J. Tsitsiklis, "Asynchronous stochastic approximation and Q-learning", *Machine Learning*, vol. 16, no. 3, pp. 185-202, 1994.

[11] C. J. C. H. Watkins, "Learning from Delayed Rewards", *Ph.D. Thesis, Uni. of Cambridge*, U.K., 1989.

[12] C. J. C. H. Watkins and P. Dayan, "Q-learning", *Machine Learning*, vol. 8, no. 3, pp. 279-292, 1992.

Fixed Points of Approximate Value Iteration and Temporal-Difference Learning*

Daniela Pucci de Farias
Management Science and Engineering
Stanford University
Stanford, CA 94305-4026
pucci@stanford.edu

Benjamin Van Roy
Management Science and Engineering
Stanford University
Stanford, CA 94305-4026
bvr@stanford.edu

Abstract

Approximate value iteration is a simple algorithm that combats the curse of dimensionality in dynamic programs by approximating iterates of the classical value iteration algorithm in a spirit reminiscent of statistical regression. Each iteration of this algorithm can be viewed as an application of a modified dynamic programming operator to the current iterate. The hope is that the iterates converge to a fixed point of this operator, which will then serve as a useful approximation of the optimal value function. In this paper, we show that, in general, the modified dynamic programming operator need not possess a fixed point, and therefore, approximate value iteration should not be expected to converge. We then propose a variant of approximate value iteration for which the associated operator *is* guaranteed to possess at least one fixed point. This variant is motivated by studies of temporal–difference learning (TD), and existence of fixed points here implies existence of stationary points for the ordinary differential equation approximated by a version of TD that incorporates "exploration."

1 Introduction

Value iteration offers a simple approach to computing optimal value functions and policies for finite–state discounted dynamic programs. The algorithm can be described compactly in terms of the "dynamic programming operator" T. In particular, value iteration generates a sequence of functions according to $J_{k+1} = TJ_k$, each mapping states to real numbers. This sequence converges to the optimal value function J^*, which is the unique fixed point of T and can be used to generate an optimal policy. However, due to the "curse of dimensionality," for problems of practical scale, the computational burden associated with storing and manipulating functions over the state space is prohibitive, and approximations are called for. One simple approximation method – dating all the way back to Bellman [1] – is approximate value iteration, which aims at approximating each iterate J_k by a linear combination of prespecified basis functions $\phi_1, \ldots, \phi_K$. Iterates $\tilde{J}_k$ are generated according to $\tilde{J}_{k+1} = \Pi T \tilde{J}_k$, where Π is a projection operator that produces a function that is in the span of $\phi_1, \ldots, \phi_K$ and close to $T\tilde{J}_k$. The hope is that $\tilde{J}_k$ converges to a good approximation of J^*.

---

*This research was supported in part by a Frederick E. Terman Award (Stanford University).

A fundamental question concerning approximate value iteration is whether the composition ΠT possesses a fixed point $\tilde{J}$ that may serve as a limit to the sequence $\tilde{J}_k$. It turns out – as will be illustrated by examples in Section 3 – that ΠT does not always have a fixed point. In subsequent sections, we propose and analyze a variant of approximate value iteration that *is* guaranteed to have a fixed point.

The variant of approximate value iteration developed in this paper was motivated by studies of temporal–difference learning (TD), a class of algorithms that can be viewed as simulation–based versions of approximate value iteration ([3], [4], [5], [6], [7], [8]). As will be discussed in our closing section, existence of fixed points for the proposed variant of approximate value iteration implies existence of stationary points for a version of TD that incorporates "exploration." Our analysis of approximate value iteration therefore also resolves an open question concerning TD.

2 Exact and Approximate Value Iteration

We consider a controlled Markov chain with a finite set of states $\mathcal{S}$ and finite sets of actions $\mathcal{A}_x, x \in \mathcal{S}$. Each state–action pair $x \in \mathcal{S}$ and $a \in \mathcal{A}_x$ is associated with a reward $g_a(x)$ and transition probabilities $P_a(x, \cdot)$. Time–relative preferences are defined by a discount factor $\alpha \in (0,1)$. We denote by P_a a matrix whose (x,y)th component is $P_a(x,y)$, and we let $P_a(x)$ be a row vector equal to the xth row of P_a.

A (stochastic stationary) policy is a mapping $\mu : \{(x,a) | \; x \in \mathcal{S}, \; a \in \mathcal{A}_x\} \mapsto [0,1]$, with $\sum_{a \in \mathcal{A}_x} \mu(x,a) = 1$ for all x. The policy defines probabilities with which actions are selected at each state. In particular, when controlled by a policy μ, the system evolves as a Markov chain with transition probabilities and costs

$$P_\mu(x,y) = \sum_{a \in \mathcal{A}_x} \mu(x,a) P_a(x,y), \; g_\mu(x) = \sum_{a \in \mathcal{A}_x} \mu(x,a) g_a(x).$$

A policy μ is deterministic if for each $x \in S$, there is an action $a \in \mathcal{A}_x$ such that $\mu(x,a) = 1$. A policy μ is optimal if it attains the supremum of

$$\mathrm{E}\left[\sum_{k=0}^{\infty} \alpha^k g_\mu(x_k) \middle| x_0 = x \right] = \left(\sum_{k=0}^{\infty} (\alpha^k P_\mu^k) g_\mu \right)(x)$$

simultaneously for all initial states $x \in \mathcal{S}$.

We assume that P_μ is irreducible and aperiodic for every μ. Hence, each P_μ possesses a unique invariant distribution π_μ with $\pi_\mu(x) > 0$ for all x. Note that π_μ is the left eigenvector associated with the unit eigenvalue of P_μ with entries adding to one.

The optimal value function J^* uniquely solves Bellman's equation $J = \max_\mu \{g_\mu + \alpha P_\mu J\}$, where the maximization is pointwise. Alternatively, defining the dynamic programming operator $TJ = \max_\mu \{g_\mu + \alpha P_\mu J\}$, the value function can be characterized as the unique fixed point of T. For each policy μ, we define an additional operator $T_\mu J = g_\mu + \alpha P_\mu J$. It is well–known that for any J, there is a deterministic policy μ such that $TJ = T_\mu J$. We call such a μ a *greedy policy* with respect to J. It is also well–known that a policy μ^* is optimal if and only if it is greedy with respect to the optimal value function J^*.

Value iteration computes improving approximations to the value function by generating a sequence according to $J_{k+1} = TJ_k$, initialized with some function J_0. It is well–known that, for any J_0, the sequence J_k converges to J^*. Unfortunately, the curse of dimensionality makes it infeasible for problems of practical scale.

Approximate value iteration aims at alleviating the prohibitive computational burden associated with value iteration by dealing with parameterized approximations rather than functions over the state space. In particular, given a preselected collection $\phi_1, \ldots, \phi_K$ of linearly independent basis functions, let $\Phi = [\phi_1 \ \cdots \ \phi_K]$,. Then the algorithm generates approximations $\tilde{J}_k = \Phi r_k$ to each iterate J_k, for some weight vector $r_k \in \Re^K$. Also, let $\phi(x) = (\phi_1(x), \ldots, \phi_K(x))'$ so that $\tilde{J}_k(x) = \phi'(x) r_k$.

The simplest form of approximate value iteration involves a projection matrix Π that projects onto the span of $\phi_1, \ldots, \phi_K$ with respect to the standard Euclidean norm, i.e., $\Pi J = \underset{\Phi r}{\operatorname{argmin}} \|J - \Phi r\|_2$. The algorithm then generates iterates according to $\tilde{J}_{k+1} = \Pi T \tilde{J}_k$.

A slightly more sophisticated version of approximate value iteration projects with respect to a weighted Euclidean norm. In particular, given a $|\mathcal{S}|$–dimensional vector π with positive components, we can define a norm by

$$\|J\|_\pi = \left(\sum_{x \in \mathcal{S}} \pi(x) J^2(x) \right)^{1/2},$$

and the associated projection is given by $\Pi J = \underset{\Phi r}{\operatorname{argmin}} \|J - \Phi r\|_\pi$. As before, approximate value iteration would generate iterates according to $\tilde{J}_{k+1} = \Pi T \tilde{J}_k$. One motivation for employing a weighted Euclidean norm is that it enables "emphasis" of "important" or "frequently visited" states in trading–off error among states.

Finally, a more general form of approximate value iteration might not employ a single projection matrix, but rather choose a projection matrix based on the current iterate. For example, given a mapping that defines for each function J a projection matrix Π_J, iterates could be generated according to $\tilde{J}_{k+1} = \Pi_{\tilde{J}_k} T \tilde{J}_k$.

3 On the Nonexistence of Fixed Points

It turns out that operators associated with simple versions of approximate value iteration often lack fixed points, hence iterates $\tilde{J}_k$ cannot converge. We illustrate potential difficulties in the next two examples, motivating the variant of approximate value iteration proposed in the next section.

In the first example, Π is the projection with respect to the standard Euclidean norm. In this example, the composition ΠT does not possess a fixed point, and approximate value iteration diverges.

Example 3.1 *Consider a Markov chain with states 1 and 2 and only one policy. In state 1, the reward is 1 and there is a probability 0.2 of remaining in that state and 0.8 of going to state 2. In state 2, the reward is 2 and there is a probability 0.2 of going to state 1 and 0.8 of staying in state 2. Let g be the vector of rewards and P be the transition probability matrix. The discount factor is $\alpha = \frac{5}{5.4}$ and we want to find an approximate value function in the subspace spanned by $\Phi = [1 \ 2]'$. Then if Π is the projection with respect to the Euclidean norm, if $\Phi r = \Pi T \Phi r$, the parameter vector r must solve $r = (\Phi'\Phi)^{-1}\Phi' (g + \alpha P \Phi r) = 1 + r$. Hence, ΠT does not have a fixed point. Furthermore, the sequence of weights r_k generated by approximate value iteration evolves according to $r_{k+1} = 1 + r_k$, and therefore the sequence is unbounded.*

The following lemma from [8] and [9] motivates the use of a certain weighted Euclidean norm in order to circumvent the difficulty that arises in the previous example.

Lemma 3.1 *Let P be the transition matrix for an irreducible aperiodic Markov chain, and let π be the invariant distribution. Then, $\|P\|_\pi \leq 1$.*

An immediate consequence of this lemma is that, for a problem with only one policy (i.e., $|\mathcal{A}_x| = 1$ for all x), the operator T is a contraction with respect to $\|.\|_\pi$, where π is the invariant distribution of the system. Furthermore, since a projection matrix Π that projects with respect to $\|\cdot\|_\pi$ is nonexpansive with respect to $\|\cdot\|_\pi$, ΠT is also a contraction. It follows that ΠT has a unique fixed point and that approximate value iteration converges to this fixed point.

The above discussion identifies a version of approximate value iteration that converges to a fixed point. Unfortunately, this only applies to problems with a single policy. Can the essential idea be generalized to problems with multiple policies? In such cases, there are usually multiple invariant distributions to take into account, each associated with a different policy. One approach to dealing with this issue involves projecting with respect to a Euclidean norm weighted by the invariant distribution associated with a policy that is greedy with respect to the current iterate $\tilde{J}_k$. In particular, let μ_J be a greedy policy with respect to J, and let for any μ, let Π_μ be a projection matrix that projects onto the span of $\phi_1, \ldots, \phi_K$ with respect to $\|\cdot\|_{\pi_\mu}$. A version of approximate value iteration can be defined by $\tilde{J}_{k+1} = \Pi_{\mu_J} TJ$. We define the operators

$$HJ = \Pi_{\mu_J} TJ \quad \text{and} \quad H_\mu J = \Pi_\mu T_\mu J.$$

To make this definition unambiguous, we need to establish what policy μ_J must be used if there is more than one greedy policy – in this case, we let μ_J be the randomized policy that takes each greedy action with the same probability.

The question that arises is whether H possesses a fixed point. We have already seen that in the case of a single policy the answer is affirmative. The following example, adapted from [3], shows that H does not necessarily possess fixed points when there are multiple policies.

Example 3.2 *Consider a controlled Markov chain with three states and two deterministic policies 1 and 2, with rewards and transition matrices given by*

$$g_1 = g_2 = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}, P_1 = \begin{bmatrix} 0.2 & 0 & 0.8 \\ 0.4 & 0.6 & 0 \\ 0 & 1 & 0 \end{bmatrix}, P_2 = \begin{bmatrix} 0.2 & 0 & 0.8 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Note that there are two possible actions at state 2, while no choices are offered at states 1 and 3. Let $\alpha = 0.99$ and $\Phi = [0\ 1\ 2]'$. For any function J, there are three possibilities for μ_J: μ_1 (with $P_{\mu_1} = P_1$), μ_2 (with $P_{\mu_2} = P_2$), and μ_3 (with $P_{\mu_3} = (P_1 + P_2)/2$).

A function J is a fixed point of H if and only if $\Pi_{\mu_J} TJ = J$, or equivalently, $\Pi_{\mu_J} T_{\mu_J} J = J$. Based on arguments made earlier regarding approximate value iteration for problems with a single policy, each composition $\Pi_{\mu_i} T_{\mu_i}$ ($i = 1, 2, 3$) has a unique fixed point. Let us denote these fixed points by $J_1^* = \Phi r_1^*$, $J_2^* = \Phi r_2^*$, and $J_3^* = \Phi r_3^*$, respectively. It turns out that $r_1^* = -0.1647$, $r_2^* = 0.3311$ and $r_3^* = 0.1889$, and that $\mu_{J_1} = \mu_2$, $\mu_{J_2} = \mu_1$, and $\mu_{J_3} = \mu_1$. It follows that neither J_1, J_2, nor J_3, are fixed points of H, and therefore H has no fixed points.

Based on the above example, one might speculate that nonexistence of fixed points may be a consequence of discontinuities of H at points where there is more than one greedy policy. As we show in the sequel, incorporating "exploration" (i.e., randomizing policies) leads to a continuous variant of H for which fixed points are guaranteed to exist.

4 Incorporating Exploration

We now introduce a modified dynamic programming operator. This definition makes use of δ–greedy policies, which effectively incorporate exploration into a greedy policy. Formally, for any $\delta > 0$, we define a δ–greedy policy μ_J^δ with respect to J by

$$\mu_J^\delta(x,a) = \frac{\exp\left[(g_a(x) + \alpha P_a(x)J)/\delta\right]}{\sum_{\bar{a}\in\mathcal{A}_x} \exp\left[(g_{\bar{a}}(x) + \alpha P_{\bar{a}}(x)J)/\delta\right]},$$

for all $x \in \mathcal{S}$ and $a \in \mathcal{A}_x$. Our modified dynamic programming operator T_δ, which we will refer to as the δ–greedy dynamic programming operator, is then defined by $T_\delta J = T_{\mu_J^\delta} J$.

Let us now establish some properties of δ–greedy policies and dynamic programming operators. Our first lemma shows that δ–greedy policies become greedy as $\delta \downarrow 0$.

Lemma 4.1 *Take* $h \in R^m$ *and let* $\eta^\delta(h,i) = \frac{\exp[h(i)/\delta]}{\sum_{j=1}^m \exp[h(j)/\delta]}$, *for* $i = 1,\dots,m$. *Then* $\sup_h \left\{\max_i h(i) - \sum_{i=1}^m \eta^\delta(h,i)h(i)\right\} \le \frac{\delta(m-1)}{e}$.

Proof: Without loss of generality, suppose that $h(m) = \max_i h(i)$. Then

$$\begin{aligned}
&\sup_h \left\{\max_i h(i) - \sum_{i=1}^m \eta^\delta(h,i)h(i)\right\} \\
&= \sup_h \sum_{i=1}^{m-1} \frac{(h(m)-h(i))\exp\left[(h(i)-h(m))/\delta\right]}{1+\sum_{j=1}^{m-1}\exp\left[(h(j)-h(m))/\delta\right]} \\
&\le \delta \sup_h \sum_{i=1}^{m-1} \frac{h(m)-h(i)}{\delta} \exp\left[(h(i)-h(m))/\delta\right] \\
&\le \delta(m-1)\sup_{x\ge 0} x\exp(-x) \\
&\le \delta(m-1)/e
\end{aligned}$$

□

As an immediate consequence of the previous lemma, we have

Lemma 4.2 $\lim_{\delta\downarrow 0}\sup_{J,x} |(T_\delta J)(x) - (TJ)(x)| = 0$.

Our next lemma establishes existence of a fixed point.

Lemma 4.3 *For any* $\delta > 0$, T_δ *has a fixed point.*

Proof: Let $G = \sup_{x\in\mathcal{S},a\in\mathcal{A}_x} |g_a(x)|$. Since there is only a finite number of states and actions, G is finite. Now consider the compact convex set $\{J : \|J\|_\infty \le G/(1-\alpha)\}$. This set is closed under T_δ, and since T_δ is continuous, Brouwer's fixed point theorem guarantees existence of a fixed point. □

Let us introduce the notion of a *quasi-contraction.*

Definition 4.1 .Quasi-contraction. *An operator* F *is a quasi-contraction with respect to a norm* $\|\cdot\|$ *if there exists a nonempty set* X^* *of fixed points, a compact set* $\mathcal{C} \supseteq X^*$, *and a scalar* $\beta \in [0,1)$ *such that for any* $x \notin \mathcal{C}$, *there exists a fixed point* $x^* \in X^*$ *such that* $\|Fx - x^*\| \le \beta\|x - x^*\|$.

Given this definition, we have the following lemma.

Lemma 4.4 *For any* $\delta > 0$, T_δ *is a quasi-contraction.*

Proof: For any J_1 and J_2,

$$\begin{aligned}\|T_\delta J_1 - T_\delta J_2\|_\infty &\leq \|TJ_1 - TJ_2\|_\infty + \|TJ_1 - T_\delta J_1\|_\infty + \\ &\quad + \|TJ_2 - T_\delta J_2\|_\infty \\ &\leq \alpha\|J_1 - J_2\|_\infty + O(\delta),\end{aligned}$$

From Lemma 4.3, T_δ has a fixed point, thus it is indeed a quasi-contraction. □

The next lemma establishes that fixed points of T_δ approximate those of T.

Lemma 4.5 *Let J^* be the unique fixed point of T, and for any $\delta > 0$, let $\mathcal{J}^\delta$ be the set of fixed points of T_δ. Then,*

$$\lim_{\delta \downarrow 0} \sup_{J \in \mathcal{J}^\delta, x \in \mathcal{S}} |J(x) - J^*(x)| = 0.$$

Proof: For any $J \in \mathcal{J}^\delta$,

$$\begin{aligned}\|J - J^*\|_\infty &= \|T_\delta J - J^*\|_\infty \\ &\leq \|T_\delta J - TJ\|_\infty + \|TJ - J^*\|_\infty \\ &\leq \alpha\|J - J^*\|_\infty + O(\delta),\end{aligned}$$

and $\|J - J^*\|_\infty = O(\delta)$. Note that the $O(\delta)$ term in the final inequality is uniformly bounded over J by Lemma 4.2. □

Lemma 4.4 has some important implications. Note that all fixed points of T_δ lie within a ball of radius $O(\delta)$. Also, for J outside this circle, T_δ behaves somewhat like a contraction. Hence, for a variant of value iteration taking the form $J_{k+1} = T_\delta J_k$, after a finite number n of iterations, iterates J_k for $k \geq n$ will all lie in this ball. Furthermore, applying Lemma 4.5, we can deduce that there is some $\bar{\delta} > 0$ such that for all $\delta \leq \bar{\delta}$, greedy policies associated with functions in the ball under consideration are optimal.

5 Existence of Fixed Points

Based on the operator T_δ, we can define a new version of approximate value iteration, which updates weights according to $\tilde{J}_{k+1} = \Pi_{\mu^\delta_{\tilde{J}_k}} T_\delta \tilde{J}_k$. Alternatively, defining

$$H_\delta J = \Pi_{\mu^\delta_J} T_\delta J,$$

we have $\tilde{J}_{k+1} = H_\delta \tilde{J}_k$. The following theorem, which is the main result of this section, establishes that, unlike H, H_δ always possesses a fixed point.

Theorem 5.1 *For any $\delta > 0$, H_δ has a fixed point.*

To aide in the proof of this theorem we will first establish a few lemmas. Henceforth we use the shorthand notation Π^δ_r to refer to $\Pi_{\mu^\delta_{\Phi r}}$ and μ^δ_r to refer to $\mu^\delta_{\Phi r}$.

We first state, without proof, continuity of certain functions.

Lemma 5.1 *The invariant distribution π_μ is a continuous function of μ.*

As discussed earlier, for each policy μ, there exists a unique vector r_μ such that $\Phi r_\mu = H_\mu \Phi r_\mu$ (this follows from Lemma 3.1). The next lemma establishes that the solution to this equation is continuous in μ.

Lemma 5.2 *The unique solution r_μ to $\Phi r_\mu = H_\mu \Phi r_\mu$ is a continuous function of μ.*

We now proceed to the main analysis. For any policy μ, let us define

$$s_\mu(r) = \Phi' D_\mu(T_\mu \Phi r - \Phi r), \; s_\delta(r) = \Phi' D_{\mu_r^\delta}(T_\delta \Phi r - \Phi r)$$

where $D_\mu = \operatorname{diag}(\pi_\mu)$ for any policy μ. We also define functions $F_\mu^\gamma : \Re^K \mapsto \Re^K$ and $F_\delta^\gamma : \Re^K \mapsto \Re^K$ by

$$F_\mu^\gamma(r) = r + \gamma s_\mu(r) \quad \text{and} \quad F_\delta^\gamma(r) = r + \gamma s_\delta(r).$$

The following lemma relates fixed points of F_μ^γ and F_δ^γ to those of H_μ and H_δ.

Lemma 5.3 *For any $\delta > 0$ and $\gamma > 0$, a vector r is a fixed point of F_μ^γ $(F^{\delta,\gamma})$ if and only if Φr is a fixed point of H_μ (H_δ).*

Proof: Let r be a fixed point of F_μ^γ. Then,

$$\begin{aligned}
s_\mu(r) &= 0 \\
\Phi' D_\mu \Phi r &= \Phi' D_\mu(g_\mu + \alpha P_\mu \Phi r) \\
\Phi(\Phi' D_\mu \Phi)^{-1} \Phi' D_\mu \Phi r &= \Phi(\Phi' D_\mu \Phi)^{-1} \Phi' D_\mu(g_\mu + \alpha P_\mu \Phi r) \\
\Phi r &= \Pi_\mu T_\mu \Phi r,
\end{aligned}$$

and Φr is a fixed point of H_μ. Reversing the steps, it can be shown that if Φr is a fixed point of H_μ, r is a fixed point of F_μ^γ. An analogous argument establishes that r is a fixed point of F_δ^γ if and only if Φr is a fixed point of H_δ. □

The next lemma establishes that, for sufficiently small γ, F_μ^γ is a pseudo–contraction.

Lemma 5.4 *There exists a constant $\gamma^* > 0$ such that for all μ and any $\gamma \in (0, \gamma^*)$, there exists a scalar $\beta_\gamma \in (0,1)$ such that $\|F_\mu^\gamma(r) - r_\mu\|_2 \leq \beta_\gamma \|r - r_\mu\|_2$.*

Proof: First, note that for all μ, $\|H_\mu \Phi r - \Phi r_\mu\|_\mu \leq \alpha \|\Phi r - \Phi r_\mu\|_\mu$, and

$$\begin{aligned}
\langle \Phi r - \Phi r_\mu, H_\mu \Phi r - \Phi r \rangle_\mu &= \langle \Phi r - \Phi r_\mu, (H_\mu \Phi r - \Phi r_\mu) + (\Phi r_\mu - \Phi r) \rangle_\mu \\
&\leq \|\Phi r - \Phi r_\mu\|_\mu \|H_\mu \Phi r - \Phi r_\mu\|_\mu - \|\Phi r - \Phi r_\mu\|_\mu^2 \\
&\leq (\alpha - 1) \|\Phi r - \Phi r_\mu\|_\mu^2 \\
&\leq (\alpha - 1)(r - r_\mu)'(\Phi' D_\mu \Phi)(r - r_\mu).
\end{aligned}$$

Since D_μ is positive definite for all μ and the set of all randomized policies is compact, there exists a constant $C_1 > 0$ independent of μ such that $(r - r_\mu)' s_\mu(r) \leq -C_1 \|r - r_\mu\|_2^2$.

Note that

$$\begin{aligned}
\|s_\mu(r)\|^2 &= \sum_{i=1}^K \left(\phi_i' D_\mu (T_\mu \Phi r - \Phi r)\right)^2 \\
&= \sum_{i=1}^K \left(\phi_i' D_\mu (\Pi_\mu T_\mu \Phi r - \Phi r)\right)^2 \\
&\leq \sum_{i=1}^K \|\phi_i\|_\mu^2 \|\Pi_\mu T_\mu \Phi r - \Phi r\|_\mu^2 \\
&\leq \sum_{i=1}^K \|\phi_i\|_\mu^2 \left(\|\Pi_\mu T_\mu \Phi r - \Phi r_\mu\|_\mu + \|\Phi r_\mu - \Phi r\|_\mu\right)^2 \\
&\leq \sum_{i=1}^K \|\phi_i\|_\mu^2 \left(\alpha \|\Phi r - \Phi r_\mu\|_\mu + \|\Phi r_\mu - \Phi r\|_\mu\right)^2 \\
&= (1 + \alpha)^2 \sum_{i=1}^K \|\phi_i\|_\mu^2 \|\Phi r_\mu - \Phi r\|_\mu^2,
\end{aligned}$$

and there exists a constant $C_2 > 0$ independent of μ such that $\|s_\mu(r)\|_2^2 \leq C_2\|r - r_\mu\|_2^2$. The inequalities we have established yield

$$\begin{aligned}\|F_\mu^\gamma(r) - r_\mu\|_2^2 &= \|r + \gamma s_\mu(r) - r_\mu\|_2^2 \\ &= \|r - r_\mu\|^2 + 2\gamma(r - r_\mu)' s_\mu(r) + \gamma^2\|s_\mu(r)\|^2 \\ &\leq (1 - 2\gamma C_1 + \gamma^2 C_2)\|r - r_\mu\|_2^2.\end{aligned}$$

The result then follows with $\gamma^* = 2C_1/C_2$. □

Lemma 5.5 *For any $\gamma > 0$ and $\delta > 0$, the function F_δ^γ possesses a fixed point.*

Proof: By Lemma 5.2, r_μ is a continuous function of μ. Since μ occupies a compact set, so does the set $R = \{r_\mu | \mu \in \gamma\}$. Let $\overline{R} = \max\{\|r\| | r \in R\}$.

Note that we only have to establish that a fixed point exists for a particular $\gamma > 0$, since, by Lemma 5.3 this fixed point is also a fixed point for all other values of γ.

Set $\gamma > 0$ such that there is a $\beta \in (0, 1)$ with $\|F_\mu^\gamma(r) - r_\mu\|_2 \leq \beta\|r - r_\mu\|_2$, for all μ. (Existence of such a γ is ensured by Lemma 5.4.) We then have

$$\begin{aligned}\|F_\delta^\gamma(r)\|_2 &\leq \|F_\delta^\gamma(r) - r_{\mu_r^\delta}\|_2 + \|r_{\mu_r^\delta}\|_2 \\ &\leq \beta\|r - r_{\mu_r^\delta}\|_2 + \overline{R} \\ &\leq \beta\|r\|_2 + (1 + \beta)\overline{R}.\end{aligned}$$

It follows that $\mathcal{C} = \left\{r \middle| \|r\|_2 \leq (1 + \beta)\overline{R}/(1 - \beta)\right\}$, is closed under F_δ^γ. The result is then a consequence of Brouwer's fixed point theorem. □

Theorem 5.1 follows from Lemmas 5.3 and 5.5.

Note that by replacing F_δ^γ with F^γ (defined in the same way as F_δ^γ with T replacing T_δ and $\mu_r = \mu_{\Phi r}$ replacing μ_r^δ), all steps in the proof of Lemma 5.5 remain valid except for the application of Brouwer's fixed point theorem, which can no longer be applied because F^γ may not be continuous because the greedy policy μ_r and the invariant distribution π_{μ_r} are not continuous in r. Nevertheless, Theorem 5.1 allows us to identify a sufficient condition for H to have a fixed point, as will be established in the upcoming theorem.

Let $\mathcal{V}_\delta$ be the set of fixed points of F_δ^γ, and let $\mathcal{P}$ be the set of vectors r such that more than one policy is greedy with respect to Φr. It is easy to show that $\mathcal{P}$ is closed. Finally, let $\mathcal{Q}_\epsilon = \{r | \|r - \hat{r}\| \geq \epsilon \text{ for all } \hat{r} \in \mathcal{P}\}$.

Theorem 5.2 *Suppose that there is $\epsilon > 0$, a decreasing sequence δ_k converging to 0, and a sequence $r_k \in \mathcal{V}_{\delta_k} \cap \mathcal{Q}_\epsilon$. Then, there exists r^* such that $\Phi r^* = H\Phi r^*$.*

Proof: First, let $a_r(x) \operatorname*{argmax}_{a \in \mathcal{A}_x} \{g_a(x) + P_a(x)\Phi r\}$ and

$$\Delta_\delta(r, x, a) = g_{a_r(x)}(x) + \alpha P_{a_r(x)}(x)\Phi r - g_a(x) - \alpha P_a(x)\Phi r.$$

Then $\inf_{r \in \mathcal{Q}_\epsilon, x, a \neq a_r(x)} \Delta_\delta(r, x, a) = \bar{\Delta} > 0$, and for all $r \in \mathcal{Q}_\epsilon$,

$$\mu_r^\delta(x, a_r(x)) \left[1 + (m - 1)\exp\left[-\bar{\Delta}/\delta\right]\right]^{-1},$$

where m is the maximum number of actions per state. Let $s(r) = \langle \phi_k, T_{\mu_r}\Phi r - \Phi r\rangle_{\mu_r}$. For small enough δ, $\|\mu_r - \mu_r^\delta\|_\infty \leq (m - 1)\exp\left[-\bar{\Delta}/\delta\right]$, thus μ_r^δ converges uniformly to

μ_r in $\mathcal{Q}_\epsilon$. Now

$$\begin{aligned}
\|s(r_k)\|_\infty &= \|s(r_k) - s_\delta(r_k)\|_\infty \\
&= \left\| \Phi' \left[D_{\mu_{r_k}} \left(g_{\mu_{r_k}} + (\alpha P_{\mu_{r_k}} - I)\Phi r_k \right) - D_{\mu^\delta_{r_k}} \left(g_{\mu^\delta_{r_k}} + (\alpha P_{\mu^\delta_{r_k}} - I)\Phi r_k \right) \right] \right\|_\infty \\
&\leq \|\Phi' D_{\mu_{r_k}} g_{\mu_{r_k}} - \Phi' D_{\mu^\delta_{r_k}} g_{\mu^\delta_{r_k}} \|_\infty + \\
&\quad + \max_i \left\| \Phi' \left[(\alpha P'_{\mu_{r_k}} - I) D_{\mu_{r_k}} - (\alpha P'_{\mu^\delta_{r_k}} - I) D_{\mu^\delta_{r_k}} \right] \phi_i \right\|_2 \frac{(1+\beta)\overline{R}}{1-\beta}.
\end{aligned}$$

Since D_μ, g_μ and P_μ are continuous functions of μ and μ^δ_r converges to μ_r, $s(r_k)$ converges to 0. Hence, H has a fixed point. □

Remark 5.1 *We have established that H_δ possesses at least one fixed point. It would be convenient if the fixed point were known to be unique, but unfortunately, it can be shown by counterexamples that this is not always the case.*

6 TD and its Stationary Points

The version of approximate value iteration that we have presented is related to and motivated by TD. The latter is a stochastic algorithm that adapts approximation weights r during simulation of the underlying Markov decision process. In this section, we will describe a version of the algorithm known as TD(0) and discuss how its stationary points coincide with fixed points of approximate value iteration. We depart, though, from the degree of rigor maintained in previous sections and only present heuristic arguments.

Application of TD(0) entails simulating a single endless trajectory x_t of the Markov decision process under consideration. The weight vector is updated upon each transition, generating a sequence r_t. Given the state x_t and decision a_t at time t, if the next state is x_{t+1}, the weight vector r_t is updated according to

$$r_{t+1} = r_t + \gamma_t \phi(x_t) \left(g_{a_t}(x_t) + \alpha(\Phi r_t)(x_{t+1}) - (\Phi r_t)(x_t) \right).$$

But how is the decision a_t selected? One simple approach is to take a greedy policy with respect to $J = \Phi r_t$. Note that such decisions are optimal if $\Phi r_t = J^*$. The hope is that, though weights may initially lead to inaccurate approximations of J^* and poor decisions, as the simulation progresses, weights will converge to those that generate accurate approximations and near–optimal decisions.

Unfortunately, the use of greedy decisions in TD(0) has appeared to perform poorly in practice (e.g., see [10]). Experiments point to the importance of "exploration;" i.e., randomization of the policy. One approach to exploration, which is connected to the variant of approximate value iteration studied in previous sections, selects decisions by letting $a_t = a$ with probability $\mu^\delta_{r_t}(x_t, a)$, for each $a \in \mathcal{A}_x$. Using results from stochastic approximation theory (e.g., see [2]), one can show that if step sizes γ_t diminish at an appropriate rate, the process followed by r_t asymptotically approximates an ordinary differential equation

$$\dot{r} = \Phi' D_{\mu^\delta_r} (g_{\mu^\delta_r} + \alpha P_{\mu^\delta_r} \Phi r - \Phi r).$$

Intuitively, this ordinary differential equation drives r in the expected direction that would be taken by the stochastic algorithm, where the expectation is taken over the steady state distribution of μ^γ_r. It is easy to show using the same reasoning as in Lemma

5.3 that a vector r is a stationary point of this ordinary differential equation if and only if it is a fixed point of our version of approximate value iteration. It follows that TD(0) with this form of exploration possesses stationary points. Note that, if greedy decisions are employed, the expected update direction when the weight vector is r is given by

$$\Phi' D_{\mu_r}(g_{\mu_r} + \alpha P_{\mu_r}\Phi r - \Phi r).$$

For this quantity to be zero, we must have $\Phi r = \Pi_{\mu_r} T \Phi r$, i.e., Φr must be a fixed point of the version of approximate value iteration considered in Example 2. As illustrated in that example, the associated operation need not possess a fixed point.

In this paper, we have established existence of fixed/stationary points for appropriate versions of approximate value iteration and TD. This is just one basic property, and a number of important questions remain open. For instance, it would be interesting to know whether the proposed versions of approximate value iteration and/or TD converge, and how well the fixed points of approximate value iteration approximate the optimal value function.

References

[1] Bellman, R., & Dreyfus, S. (1959). Functional approximations and dynamic programming. *Mathematical Tables and Other Aids to Computation*, *13*, 247-251.

[2] Benveniste, A. Métivier M., & Priouret, P. (1990). *Adaptive algorithms and stochastic approximation.* Berlin: Springer-Verlag.

[3] Bertsekas, D. P., & Tsitsiklis, J. N. (1995). *Neuro-dynamic programming.* Belmont, MA: Athena Scientific.

[4] Dayan, P. D. (1992). The convergence of TD(λ) for general λ, *Machine Learning*, *8*, 341-362.

[5] Gurvits, L. Lin, L. J., & Hanson, S. J. (1994). *Incremental learning of evaluation functions for absorbing Markov chains: New methods and theorems.* Unpublished manuscript, Siemens Corporate Research, Princeton, NJ.

[6] Pineda, F. (1997). Mean-field analysis for batched TD(λ). *Neural Computation*, *9*, 1403–1419.

[7] Sutton, R. S. (1988). Learning to predict by the method of temporal differences. *Machine Learning*, *3*, 9-44.

[8] Tsitsiklis, J. N. & Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, *42*, 674-690.

[9] Van Roy, B. (1998). *Learning and value function approximation in complex decision processes.* Ph.D. dissertation, Department of Electrical Engineering and Computer Science, MIT, Cambridge, MA.

[10] Van Roy, B. Bertsekas, D. P. Lee, Y., & Tsitsiklis, J. N. (1997). A neuro-fynamic programming approach to retailer inventory management. *Proceedings of the IEEE Conference on Decision and Control.*

Some Applications of Reinforcement Learning To Real-World Problems

Abhijit Gosavi

Dept. of Engineering
University of Southern Colorado
261, Technology Building, 2200 Bonforte Blvd, Pueblo, CO 81001
gosavi@uscolo.edu

Abstract

Reinforcement Learning is a simulation-based version of solving Markov Decision Problems (MDPs) and Semi-Markov Decision Problems (SMDPs), using Bellman's dynamic programming (DP) framework. It is especially useful in solving large-scale MDPs and SMDPs, where the twin curses of dimensionality and modeling make it difficult to apply classical DP techniques such as value and policy iteration.

I shall present an overview of three case-studies from the real world, where the problems are too large and the transition mechanisms too complex for one to be able to accurately calculate or even store the associated transition matrices. On such problems, classical DP is not very useful and reinforcement learning is a possible alternative. The problems, I shall discuss, are related to airline yield-management, supply chain management and maintenance management. I shall also enumerate some computational issues with roots in the theory of function approximation, which form an important element of getting these methods to work.

1 Introduction

Problems of sequential decision-making under uncertainty, whose stochastic structures have an underlying Markov chain, can often be modeled as Markov Decision Problems (MDPs) or its variants. Bellman's equation [1] forms the key to solving these problems. Stochastic dynamic programming (DP) methods such as value and policy iteration, which use the Bellman equation, can be employed for this purpose [16]. However, classical versions of these methods break down in the face of 1) a large state-space (*curse of dimensionality*) and 2) a complex stochastic structure (*curse of modeling*). In the presence of a large state-space with millions of states, it is difficult to store or compute the elements of the transition probability matrices, which is an integral element of value or policy iteration. In addition, it is difficult to set up the linear equations needed in policy iteration or compute the elements of the value function needed in value iteration. Even for small state-spaces, where storage may not be a problem, often a large number of random variables in the stochastic system make it difficult to calculate the transition probabilities of the underlying Markov chain. Under such conditions, it is almost impossible to use the Bellman equation directly. As a result, recent years have seen the stimulation of much research interest in alternatives to solving these problems. The theory of learning automata [11] forms a simulation-based, game-theoretic framework for solving these problems. Reinforcement Learning (RL) is yet another alternative that promises to alleviate DP of its curses. RL is also simulation-based and is essentially a version of the Bellman equation based methods.

Although a very large volume of literature in operations research is devoted to the solution methods and applications of the MDP, commercial applications to real world problems have been relatively rare. Of course the twin curses, mentioned above, are the chief culprits because most real world problems tend to have a large state-space and often a complex stochastic structure. It was only recently that an MDP-based model that used RL was employed in a commercial elevator scheduling problem [5] where the state-space was of the order of millions of states. RL, naturally, has opened exciting possibilities of tapping the power of the MDP in real-world applications. As a result, analysis of its convergence properties assumes great significance. A number of papers dealing with these issues have been written. See Tsitsiklis [18], Borkar [3] and Borkar and Meyn [4]. See also Bertsekas and Tsitsiklis [2]. A careful study of these analyses is crucial in obtaining success in computational work. Hence I shall first enumerate a couple of theoretical issues related to convergence that play an important role in computational experiments. Thereafter I shall discuss three case-studies, where RL can be applied. The rest of this paper is organized as follows. Section 2 discusses in brief the use of neural networks in RL. In Section 3, I discuss briefly an issue related to the convergence of model-free RL. Sections 4, 5 and 6, discuss RL case-studies related to airline yield-management, preventive maintenance and supply-chain inventory control, respectively. Section 7 concludes the paper.

2 Use of Neural Networks in RL

To approximate the elements of the value function, a backpropagation- controlled neural network can be used. The idea is to store the large number of elements of the value function, which in RL is computed in the form of the so-called Q-factors, in the form of a few scalars (also known as the weights of the neural network). A major problem with this approach is that the value function keeps changing in any dynamic programming scheme. As a result, combining a neural network based on backpropagation with an RL algorithm that updates the Q-factors needs to be analyzed for its convergence. Unfortunately, there are no known results that guarantee its convergence under the assumption of a backpropagation neural network. Not surprisingly, many instances of divergence with neural networks have been reported in literature. There are, however, theoretical results [2] that show convergence if states are aggregated to form super-states. Computational work done by this author also points towards the robustness of this approach (see [12]). To summarize, although neural networks have been widely advocated, their robustness and convergence properties have been attacked both from the theoretical and practical standpoints. We, next, discuss issues related to solution methods that do not generate the transition probability matrix.

3 Model-Free RL and Step Size

RL methods that seek to generate solutions without computing the transition probability matrix are called *model-free* methods. The idea is due to Robbins and Monro [17], who first suggested a stochastic approximation scheme for approximating the value of a random variable using simulation, without the knowledge of the probability density function of the random variable. The scheme is of the form $v \leftarrow (1-\mu)v + \mu a$, where v is an estimate of the random variable, a is a sample generated from the simulator, and μ is the step-size or learning rate. They showed that a step-size equal to $1/k$, where k denotes the iteration number, would generate the expected value of the random variable as $k \to \infty$. Q-learning uses the same scheme to generate the Q-factor from its samples (the expression for the sample is derived from the Bellman equation), without the knowledge of the transition probabilities. Accordingly, for a given state-action pair (i, a), the Q-factor denoted by $Q(i, a)$ should use a step-size equal to $1/k(i, a)$, if $k(i, a)$ were to denote the number of trials of the state-action pair (i, a) in the simulator. However, this raises an important question. Suppose there are a thousand state-action pairs in a problem. Then one would have to store a thousand scalars - one for each $k(i, a)$. For a large state-space (of the order of a billion states), this approach will run into trouble because storing the k terms in a function approximation scheme is ruled out (this is because they increase continuously). This, once again, suggests that combining states into a small number of super-states is a more tractable approach, where one can not only store the Q-factors for the super-states explicitly, but also store the $k(i, a)$ terms explicitly. In summary, the selection of the proper-step size is critical for the success of RL and may dictate the nature of the function approximation scheme. We, next, discuss a case-study related to airline yield management.

4 Use of RL in Airline Yield Management

The deregulation of the airline industry in 1978 allowed airline companies to choose their own market segments and routes and set their own fares. With the fierce competition that ensued in the airline industry as a result of this, it was soon realized that optimization techniques were needed for running their operations. An important goal of any airline company is that of maximizing the revenue obtained from the sale of tickets on any given flight. This is referred to in the literature as *yield management.* It forms a crucial aspect of airline logistics. According to a report from SIAM news [8]: "Yield management ... saved American Airlines \$1.4 billion in the period from 1989 to 1992, about 50% more than its net profit of \$892 million for the same period. Modeling and optimization made the difference between profit and loss." And according to a very recent report [13] "By 1998,...the revenue management system at American Airlines was generating nearly \$1 billion in annual incremental revenue. To put that figure into perspective, consider that the airline's total operating profit didn't approach \$1 billion till 1997."

Seat allocation and overbooking are two critical aspects of the yield management problem and the key to making profit lies in exercising control over the seat allocation and the overbooking aspects. The problem considered in Gosavi [12] is a combined problem of seat allocation and overbooking having the following features: a single flight leg with multiple fare classes, concurrent and random demand arrivals from the different fare classes, class-dependent and random cancellations.

Accommodating all these features provides a computational challenge because that makes it a large-scale and complex stochastic optimization problem. Most models in the existing literature accommodate only a subset of these features to make the model tractable. At the same time, because of the high stakes involved, there is a need for an approach that accounts for all these factors and generates optimal or near-optimal results. In fact, according to a recent review paper on revenue management written by McGill and Ryzin [14]: "Recently developed approximation methods for dynamic programming and stochastic programming (reinforcement learning) may be useful in revenue management." Hence RL obviously seems to have the potential for offering solutions here. Gosavi [12] uses a policy iteration based approach on this problem taking into consideration all the features mentioned above. For a value iteration based RL approach to this problem, see Gosavi *et al* [10]. In both papers, a well-known heuristic has been outperformed.

The seat allocation problem results from the airlines' practice of selling seats within the same cabin of a flight at different prices. There are differences in customers and airlines exploit these differences to sell tickets at different prices. A passenger who desires a lesser number of stopovers or a passenger who arrives late in the booking process is made to pay a higher fare. And this results in classifying passengers into different fare-classes based on their needs and the circumstances. Airlines protect some seats from the lower revenue fare classes in order to be able to satisfy demands from the higher revenue classes since that helps in increasing the revenue. But then the obvious question that arises is: what is the optimal level of protection ?

The 'overbooking' aspect adds another complicating dimension to this problem. It arises from the fact that customers do cancel their tickets and often fail to show up at the flight time (no-shows). Airlines tend to "overbook" fights (sell more tickets than the number of seats available) anticipating such events to avoid flying with empty seats. However, excessive overbooking carries the risk of not having enough seats for all the ticket-holders that show up. When this happens, airlines are forced to deny (bump) boarding request to the extra ticket- holders and pay a penalty in two ways: directly in the form of financial compensations to the inconvenienced passengers, and indirectly through loss of customer goodwill. Hence, a careful choice of overbooking policy that maximizes the yield is required.

In order to model the yield-management problem as a semi-Markov decision problem (SMDP), we first define the system state-space. Conservatively, for n fare classes, the system state (ϕ) can be denoted by the vector

$$\phi = (c, s_1, s_2, ..., s_n, \psi_1, \psi_2, ..., \psi_n, t),$$

where c represents the class of the most recent customer (among n possible classes) seeking a ticket, $s_1, s_2, ..., s_n$ represent the number of seats sold in n classes, ψ_k, for $k \in \{1, 2, ..., n\}$, is a vector of size s_k containing the times of arrivals of all the customers with a class k ticket, and t represents the time remaining to departure of the flight. The cardinality of the finite part of the state space can be shown to have an upper bound of $n * M^n$, where M is the maximum number of customers in any given class with tickets for the plane. Note that for all practical purposes, M could be equal to the total flight capacity plus the overbooking amount.

Clearly, a change in the system state is triggered by any one of the following three events: 1) a new customer arrives to the system requesting a ticket, 2) a cancellation occurs and 3) the flight departs. Whenever a customer places a request for a seat, a decision has to be made regarding whether to accept or deny the request. The action space in this problem is $A = \{accept, deny\}$. It may be noted that the decision-maker can address both the seat-allocation problem and the overbooking aspects of the yield management problem by selecting a policy vector (π) of the size of the state-space that contains either "accept" or "deny" decision for every state visited by the system. The metric used in Gosavi [12] is long-run average reward. For more details and actual results, the reader is referred to [12]. We next present a case-study related to preventive maintenance where RL is used.

5 Use of RL in Preventive Maintenance

Consider a production-inventory system that consists of 1) a single machine and 2) a finished product buffer into which the parts produced by the machine go. A random demand arrival process depletes this buffer. The demand comes from customers. Satisfying demand results in profits. Any demand that arrives when the buffer is empty is lost (never to return). (A lost demand can also be viewed as a loss and taking care of this assumption is a trivial matter in the simulator.) The machine under consideration is prone to random failures. Repairs are costlier than preventive maintenance (PM). We shall assume that the machine is as good as new when it is repaired or maintained. Since the time between failures is a random variable (with known distribution), the problem here is to find when to do PM. If one waits too long the machine is likely to fail and if one maintains the machine too early, excessive maintenance costs are likely to be incurred. The time of producing one unit, the time for repair and the time for maintenance are random variables. The buffer size plays an important role in the profit dynamics of this system. The machine is assumed to operate on an (S, s) policy. When the buffer hits a size of S, the machine stops and takes a vacation till the buffer goes down to size s. The age of the machine can be measured in terms of the units produced since its last repair or maintenance, also called the *production count*. The optimization problem is hence to find the threshold production count when the machine should be maintained to maximize the net profits. The problem can be cast as an SMDP with the state of the system given by (w, c), where w denotes the age of the machine in terms of the number of parts produced since the last repair or maintenance and c denotes the number of parts in the buffer. The action space for this problem consists of two actions: 1) stop the machine for PM and 2) continue with production. At the end of producing one unit, the decision-maker has to make this decision depending on the state of the system. For more details on this problem the reader is referred to Das *et al* [6] and Gosavi [9]. The case-study comes from a paper written by Das and Sarkar [7] , where a number of modeling assumptions were made to keep the model tractable so as to obtain a closed form for the objective function. Most of these assumptions can be easily relaxed in a simulator and this is where the power of RL sets in.

6 Use of RL in Inventory Control in a Large Supply Chain

The following case-study comes from [15]. The problem is one related to inventory control in a company that has manufacturing plants and markets in several countries. Demand in each market is satisfied by a warehouse, which will be referred to as the import warehouse. Each production plant supplies material to an export warehouse. When an import warehouse runs out of material, it has to place an order to one of the different export warehouses. The cost of production and the transportation costs vary from country to country. Hence clearly, there is a need to make a decision related to which export warehouse to choose from. Orders from a warehouse in a different country may take more time and may have high transportation costs but may actually be cheaper. The cost of not being able to supply demand on time will vary from country to country. The system state is modeled as follows: $(W_1, W_2, \ldots, W_n, w_1, w_2, \ldots, w_n)$, where W_i denotes the number of items in the ith export warehouse, while w_i denotes the number of items in the ith import warehouse. The action to be taken belongs to the set $A = \{a_1, a_2 \ldots, a_n\}$, where a_i denotes the selection of the export warehouse of the ith country for placing orders. It is clear that the state-space of this problem is enormous and it is a stochastic system since the demand arrivals and transportation times are random. The problem is modeled as a semi-Markov decision problem and a value iteration based average reward Q-learning algorithm has been applied for solution purposes. It has been shown to outperform some known heuristics in the literature. For more details, the reader is referred to [15].

7 Conclusions

It is clear that RL seems to have a great potential to solve large-scale and complex MDPs and SMDPs, which were considered intractable in the past. A major advantage of RL is that a large number of realistic assumptions can be made in the modeling, which is due to the fact that the model is simulation-based. To be able to implement RL algorithms, one has to pay careful attention to their convergence properties as they frequently make the difference between success and failure. Function approximation of the value function is a research topic by itself and a lot is still unknown about the convergence behavior of RL algorithms when coupled with function approximation schemes. The success of any RL implementation when applied to a large-scale problem relies critically on the function approximation scheme.

References

[1] R. Bellman, "The theory of dynamic programming," *Bull. Amer. Math. Soc*, vol. 60, pp. 503-516,1954.

[2] D. Bertsekas and J. Tsitsiklis. *Neurodynamic Programming.* Athena Scientific, Belmont, Mass, USA, 1996.

[3] V.S. Borkar. "Asynchronous stochastic approximation," *SIAM J. Control and Optim.*, vol. 36, pp. 840-851, 1998 (Correction note in *ibid.* 39, to appear).

[4] V. S. Borkar and S.P. Meyn, "The ODE Method for Convergence of Stochastic Approximation and Reinforcement Learning," *Working paper.*

[5] R. Crites and A. Barto, "Improving Elevator Performance Using Reinforcement Learning," *Neural Information Processing Systems: Proceedings of the 1995 Conference*, pp. 1017-1023, 1996.

[6] T.K. Das, A. Gosavi, S. Mahadevan, and N. Marchalleck. "Solving Semi-Markov Decision Problems Using Average Reward Reinforcement Learning," *Management Science* April, 45(4), pp 560-574, 1999.

[7] T.K. Das and S. Sarkar, "Optimal Preventive Maintenance in Production Inventory Systems," *IIE Transactions on Quality and Reliability*, vol. 31, no. 6, pp. 537-551, 1999.

[8] P. Davis, "Airline Ties Profitability Yield to Management,"*SIAM News*, vol. 27 no. 5, 1994.

[9] A. Gosavi, "A Convergent Algorithm for Solving Semi-Markov Decision Problems Using Under Long-Run Average Cost", Under review with the *European Journal of Operational Research.*

[10] A. Gosavi, N. Bandla, and T.K. Das, "A Reinforcement Learning Approach to Airline Seat Allocation for Multiple Fare Classes with Overbooking," submitted to *IIE Transactions.*

[11] A. Gosavi, T.K. Das, and S. Sarkar, "A Simulation-Based Learning Automata Framework for Solving Semi-Markov Decision Problems Under Long-run Average Cost," Second iteration *IIE Transactions.*

[12] A. Gosavi, "A Policy Iteration Based Average Reward Reinforcement Learning Algorithm for Solving Semi-Markov Decision Problems: Convergence Analysis and Numerical Results," Under review with the *Machine Learning.*

[13] P. Horner,"Mother, Father of Invention Produce Golden Child: Revenue Management - A Sabre Story," *ORMS Today* vol. 27, no. 3, June, 2000.

[14] J.I. McGill and Garrett J. Van Ryzin, "Revenue Management: Research Overview and Prospects," *Transportation Science*, vol 33, no. 2,pp. 233-256, 1999.

[15] P. Pontrandolfo, A. Gosavi, G. Okogbaa, and T.K. Das, "Global Supply Chain Management: A Reinforcement Learning Approach," *Under review with the International Journal of Production Research.*

[16] M.L. Puterman, *Markov Decision Processes.* John Wiley, New York, 1994.

[17] H. Robbins and S. Monro, "A Stochastic Approximation Method," *Annals of Math. Statist.* vol. 22, pp 400-407, 1951.

[18] J. Tsitsiklis, "Asynchronous Stochastic Approximation and Q-Learning," *Machine Learning* vol 16, pp. 185-202, 1994.

Enhanced Optimality Conditions and Exact Penalty Functions[1]

by

Dimitri P. Bertsekas and Asuman E. Koksal[2]

Abstract

We consider optimization problems with equality, inequality, and abstract set constraints, and we explore various characteristics of the constraint set that imply the existence of Lagrange multipliers. We prove a generalized version of the Fritz-John theorem, and we introduce new and general conditions that extend and unify the major constraint qualifications. Among these conditions, a new property, pseudonormality, provides the connecting link between the classical constraint qualifications and the use of exact penalty functions.

1. INTRODUCTION

We consider finite-dimensional optimization problems of the form

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in C, \end{aligned} \tag{1.1}$$

where the constraint set C consists of equality and inequality constraints as well as an additional abstract set constraint $x \in X$:

$$C = X \cap \{x \mid h_1(x) = 0, \ldots, h_m(x) = 0\} \cap \{x \mid g_1(x) \leq 0, \ldots, g_r(x) \leq 0\}. \tag{1.2}$$

We assume throughout the paper that f, h_i, g_j are continuously differentiable functions from $\Re^n$ to $\Re$, and X is a nonempty closed set. In our notation, all vectors are viewed as column vectors, and a prime denotes transposition.

Necessary conditions for the above problem can be expressed in terms of tangent cones, normal cones, and their polars. In our terminology, a vector y is a *tangent* of a set $S \subset \Re^n$ at a vector $x \in S$ if either $y = 0$ or there exists a sequence $\{x^k\} \subset S$ such that $x^k \neq x$ for all k and

$$x^k \to x, \qquad \frac{x^k - x}{\|x^k - x\|} \to \frac{y}{\|y\|}.$$

An equivalent definition often found in the literature (e.g., [BSS93], [RoW98]) is that there exists a sequence $\{x^k\} \subset S$ with $x^k \to x$, and a positive sequence $\{\alpha_k\}$ such that $\alpha^k \to 0$ and $(x^k - x)/\alpha^k \to$

---

[1] Research supported by NSF under Grant ACI-9873339.

[2] Dept. of EE and CS, M.I.T., Cambridge, MA, 02139, dimitrib@.mit.edu, asuman@mit.edu.

y. The set of all tangents of S at x is denoted by $T_S(x)$ and is also referred to as the *tangent cone* of S at x. The *polar cone* of any cone T is defined by

$$T^\perp = \{z \mid z'y \le 0,\ y \in T\}.$$

For a nonempty cone T, we will use the well-known relation $T \subset (T^\perp)^\perp$, which holds with equality if T is closed and convex.

For a closed set X and a point $x \in X$, we will also use the *normal cone* of X at x, denoted $N_X(x)$, which is obtained from the polar cone $T_X(x)^\perp$ by means of a closure operation. In particular, we have $z \in N_X(x)$ if there exist sequences $\{x^k\} \subset X$ and $\{z^k\}$ such that $x^k \to x$, $z^k \to z$, and $z^k \in T_X(x^k)^\perp$ for all k. Equivalently, the graph of $N_X(\cdot)$, viewed as a point-to-set mapping, $\{(x,z) \mid z \in N_X(x)\}$, is the closure of the graph of $T_X(\cdot)^\perp$. The normal cone, introduced by Mordukhovich [Mor76], has been studied by several authors, and is of central importance in nonsmooth analysis (see the books by Aubin and Frankowska [AuF90], Rockafellar and Wets [RoW98], and Borwein and Lewis [BoL00]). In general, we have $T_X(x)^\perp \subset N_X(x)$ for any $x \in X$. In the case where $T_X(x)^\perp = N_X(x)$, we will say that X is *regular* at x. The term "regular at x in the sense of Clarke" is also used in the literature (see, [RoW98], p. 199). Two properties of regularity that are important for our purposes are that (1) if X is convex, then it is regular at each $x \in X$, and (2) if X is regular at an $x \in X$, then $T_X(x)$ is convex ([RoW98], pp. 203 and 221).

A classical necessary condition for a vector $x^* \in C$ to be a local minimum of f over C is

$$\nabla f(x^*)'y \ge 0, \qquad \forall\, y \in T_C(x^*), \tag{1.3}$$

where $T_C(x^*)$ is the tangent cone of C at x^* (see e.g., [BSS93], [Ber99], p. 335, [Hes75], [Roc93], [RoW98]). Necessary conditions that involve Lagrange multipliers relate to the specific representation of the constraint set C in terms of the constraint functions h_i and g_j. In particular, we say that the constraint set C of Eq. (1.2) *admits Lagrange multipliers* at a point $x^* \in C$ if for every continuously differentiable cost function f for which x^* is a local minimum of problem (1.1) there exist vectors $\lambda^* = (\lambda_1^*, \ldots, \lambda_m^*)$ and $\mu^* = (\mu_1^*, \ldots, \mu_r^*)$ that satisfy the following conditions:

$$\left(\nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla h_i(x^*) + \sum_{j=1}^r \mu_j^* \nabla g_j(x^*)\right)' y \ge 0, \qquad \forall\, y \in T_X(x^*), \tag{1.4}$$

$$\mu_j^* \ge 0, \qquad \forall\, j = 1, \ldots, r, \tag{1.5}$$

$$\mu_j^* = 0, \qquad \forall\, j \notin A(x^*), \tag{1.6}$$

where $A(x^*) = \{j \mid g_j(x^*) = 0\}$ is the index set of inequality constraints that are active at x^*. We refer to such a pair (λ^*, μ^*) as a *Lagrange multiplier vector* or simply a *Lagrange multiplier*.

Conditions that guarantee the existence of Lagrange multipliers are called *constraint qualifications*, and have been investigated extensively in the literature. Some of the most useful ones are the following:

CQ1: $X = \Re^n$ and x^* is a regular point in the sense that the equality constraint gradients $\nabla h_i(x^*)$, $i = 1, \ldots, m$, and the active inequality constraint gradients $\nabla g_j(x^*)$, $j \in A(x^*)$, are linearly independent.

CQ2: $X = \Re^n$, the equality constraint gradients $\nabla h_i(x^*)$, $i = 1, \dots, m$, are linearly independent, and there exists a $y \in \Re^n$ such that

$$\nabla h_i(x^*)'y = 0, \quad i = 1, \dots, m, \qquad \nabla g_j(x^*)'y < 0, \quad \forall\, j \in A(x^*).$$

For the case where there are no equality constraints, this is known as the Arrow-Hurwitz-Uzawa constraint qualification, introduced in [AHU61]. In the more general case where there are equality constraints, it is known as the Mangasarian-Fromovitz constraint qualification, introduced in [MaF67].

CQ3: $X = \Re^n$, the functions h_i are linear and the functions g_j are concave.

It is well-known that all of the above constraint qualifications imply that the constraint set admits Lagrange multipliers (see e.g., [Ber99] or [BSS93]). It is also well-known that constraint qualifications and Lagrange multipliers are related to exact penalty functions. In particular, let us say that the constraint set C *admits an exact penalty* at the feasible point x^* if for every continuously differentiable function f for which x^* is a strict local minimum of f over C, there is a scalar $c > 0$ such that x^* is also a local minimum of the function

$$F_c(x) = f(x) + c\left(\sum_{i=1}^{m} |h_i(x)| + \sum_{j=1}^{r} g_j^+(x)\right)$$

over $x \in X$, where we denote

$$g_j^+(x) = \max\{0, g_j(x)\}.$$

We have the following:

(a) If X is convex and the constraint set admits an exact penalty at x^* it also admits Lagrange multipliers at x^* (this follows from Prop. 3.112 of Bonnans and Shapiro [BoS00]; see also the subsequent Prop. 5).

(b) Each of the above constraint qualifications CQ1-CQ3 implies that C admits an exact penalty (the case of CQ1 was treated by Pietrzykowski [Pie69]; the case of CQ2 was treated by Zangwill [Zan67], Han and Mangasarian [HaM79], and Bazaraa and Goode [BaG82]; the case of CQ3 will be dealt with in the present paper – see the subsequent Props. 2 and 4).

In this paper, we establish the connections between constraint qualifications, Lagrange multipliers, and exact penalty functions. Much of our analysis is motivated by an enhanced set of Fritz John necessary conditions which are introduced in the next section. These conditions were proved in a somewhat weaker form for the case where $X = \Re^n$ in a largely overlooked analysis by Hestenes [Hes75] (see the discussion in Section 2). They were generalized for the case where X is a general closed convex set in the first author's recent textbook [Ber99] (Prop. 3.3.11), and they will be further generalized in Section 2 for the case where X is a closed but not necessarily convex set.

In Section 3, we introduce the new notion of *constraint pseudonormality*, and we discuss its connection with classical results relating constraint qualifications and the admittance of Lagrange multipliers. We also give a new and natural extension of the Mangasarian-Fromovitz constraint qualification that applies to the case where $X \neq \Re^n$. Finally, in Section 4, we make the connection between pseudonormality and exact penalty functions. In particular, we show that pseudonormality implies the admittance of an exact penalty, while being implied by the major constraint qualifications. In the process we prove in a unified way that the constraint set admits an exact penalty for a much larger variety of constraint qualifications than has been known hitherto.

ENHANCED FRITZ JOHN CONDITIONS

The Fritz John necessary optimality conditions [Joh48] are often used as the starting point for the analysis of Lagrange multipliers. Unfortunately, these conditions in their classical form are insufficient to derive the existence of Lagrange multipliers under some of the standard constraint qualifications, such as linearity of the constraint functions h and g. Recently, the classical Fritz John conditions have been strengthened through the addition of an extra necessary condition, and their effectiveness has been significantly enhanced (see [Hes75] for the case $X = \Re^n$, and [Ber99] for the case where X is a closed convex set). A further extension is given by the following proposition.

Proposition 1: Let x^* be a local minimum of problem (1.1)-(1.2). Then there exist scalars μ_0^*, $\lambda_1^*, \ldots, \lambda_m^*$, and $\mu_1^*, \ldots, \mu_r^*$, satisfying the following conditions:

(i) $-\left(\mu_0^* \nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla h_i(x^*) + \sum_{j=1}^r \mu_j^* \nabla g_j(x^*)\right) \in N_X(x^*)$.

(ii) $\mu_j^* \geq 0$ for all $j = 0, 1, \ldots, r$.

(iii) $\mu_0^*, \lambda_1^*, \ldots, \lambda_m^*, \mu_1^*, \ldots, \mu_r^*$ are not all equal to 0.

(iv) If the set $I \cap J$ is nonempty, where $I = \{i \mid \lambda_i^* \neq 0\}$ and $J = \{j \mid \mu_j^* > 0\}$, then given any neighborhood B of x^* and any $\epsilon > 0$, there is an $x \in B \cap X$ such that

$$f(x) < f(x^*), \quad \lambda_i^* h_i(x) > 0, \quad \forall\, i \in I, \qquad g_j(x) > 0, \quad \forall\, j \in J, \tag{2.1}$$

$$|h_i(x)| \leq \epsilon\, m(x), \quad \forall\, i \notin I, \qquad g_j(x) \leq \epsilon\, m(x), \quad \forall\, j \notin J, \tag{2.2}$$

where

$$m(x) = \min\bigl\{\min\{|h_i(x)| \mid i \in I\}, \min\{g_j(x) \mid j \in J\}\bigr\}. \tag{2.3}$$

Proof: We use a quadratic penalty function approach. For each $k = 1, 2, \ldots$, consider the "penalized" problem

$$\text{minimize } F^k(x) \equiv f(x) + \frac{k}{2}\sum_{i=1}^m \bigl(h_i(x)\bigr)^2 + \frac{k}{2}\sum_{j=1}^r \bigl(g_j^+(x)\bigr)^2 + \frac{1}{2}||x - x^*||^2$$

$$\text{subject to } x \in X \cap S,$$

where $S = \{x \mid ||x - x^*|| \leq \epsilon\}$, and $\epsilon > 0$ is such that $f(x^*) \leq f(x)$ for all feasible x with $x \in S$. Since $X \cap S$ is compact, by Weierstrass' theorem, we can select an optimal solution x^k of the above problem. We have for all k

$$f(x^k) + \frac{k}{2}\sum_{i=1}^m \bigl(h_i(x^k)\bigr)^2 + \frac{k}{2}\sum_{j=1}^r \bigl(g_j^+(x^k)\bigr)^2 + \frac{1}{2}||x^k - x^*||^2 = F^k(x^k) \leq F^k(x^*) = f(x^*) \tag{2.4}$$

and since $f(x^k)$ is bounded over $X \cap S$, we obtain

$$\lim_{k\to\infty} ||h_i(x^k)|| = 0, \quad i = 1, \ldots, m, \qquad \lim_{k\to\infty} ||g_j^+(x^k)|| = 0, \quad j = 1, \ldots, r;$$

otherwise the left-hand side of Eq. (2.4) would become unbounded from above as $k \to \infty$. Therefore, every limit point $\overline{x}$ of $\{x^k\}$ is feasible, i.e., $\overline{x} \in C$. Furthermore, Eq. (2.4) yields $f(x^k) + (1/2)||x^k - x^*||^2 \leq f(x^*)$ for all k, so by taking the limit as $k \to \infty$, we obtain

$$f(\overline{x}) + \frac{1}{2}||\overline{x} - x^*||^2 \leq f(x^*).$$

Since $\overline{x} \in S$ and $\overline{x}$ is feasible, we have $f(x^*) \leq f(\overline{x})$, which when combined with the preceding inequality yields $||\overline{x} - x^*|| = 0$ so that $\overline{x} = x^*$. Thus the sequence $\{x^k\}$ converges to x^*, and it follows that x^k is an interior point of the closed sphere S for all k greater than some $\overline{k}$.

For $k \geq \overline{k}$, we have by the necessary condition (1.3), $\nabla F^k(x^k)'y \geq 0$ for all $y \in T_X(x^k)$, or equivalently $-\nabla F^k(x^k) \in T_X(x^k)^\perp$, which is written as

$$-\left(\nabla f(x^k) + \sum_{i=1}^m \xi_i^k \nabla h_i(x^k) + \sum_{j=1}^r \zeta_j^k \nabla g_j(x^k) + (x^k - x^*)\right) \in T_X(x^k)^\perp, \tag{2.5}$$

where

$$\xi_i^k = kh_i(x^k), \qquad \zeta_j^k = kg_j^+(x^k). \tag{2.6}$$

Denote,

$$\delta^k = \sqrt{1 + \sum_{i=1}^m (\xi_i^k)^2 + \sum_{j=1}^r (\zeta_j^k)^2}, \tag{2.7}$$

$$\mu_0^k = \frac{1}{\delta^k}, \qquad \lambda_i^k = \frac{\xi_i^k}{\delta^k}, \quad i = 1, \ldots, m, \qquad \mu_j^k = \frac{\zeta_j^k}{\delta^k}, \quad j = 1, \ldots, r. \tag{2.8}$$

Then by dividing Eq. (2.5) with δ^k, we obtain

$$-\left(\mu_0^k \nabla f(x^k) + \sum_{i=1}^m \lambda_i^k \nabla h_i(x^k) + \sum_{j=1}^r \mu_j^k \nabla g_j(x^k) + \frac{1}{\delta^k}(x^k - x^*)\right) \in T_X(x^k)^\perp. \tag{2.9}$$

Since by construction we have

$$(\mu_0^k)^2 + \sum_{i=1}^m (\lambda_i^k)^2 + \sum_{j=1}^r (\mu_j^k)^2 = 1, \tag{2.10}$$

the sequence $\{\mu_0^k, \lambda_1^k, \ldots, \lambda_m^k, \mu_1^k, \ldots, \mu_r^k\}$ is bounded and must contain a subsequence that converges to some limit $\{\mu_0^*, \lambda_1^*, \ldots, \lambda_m^*, \mu_1^*, \ldots, \mu_r^*\}$.

From Eq. (2.9) and the definition of the normal cone $N_X(x^*)$, we see that μ_0^*, λ_i^*, and μ_j^* must satisfy condition (i). From Eqs. (2.6) and (2.8), μ_0^*, λ_i^*, and μ_j^* must satisfy condition (ii), and from Eq. (2.10), they must satisfy condition (iii). Finally, to show that condition (iv) is satisfied, assume that $I \cup J$ is nonempty (otherwise we are done), and note that for all sufficiently large k within the index set $\mathcal{K}$ of the convergent subsequence, we must have $\lambda_i^* \lambda_i^k > 0$ for all $i \in I$ and $\mu_j^* \mu_j^k > 0$ for all $j \in J$. Therefore, for these k, from Eqs. (2.6) and (2.8), we must have $\lambda_i^* h_i(x^k) > 0$ for all $i \in I$ and $\mu_j^* g_j(x^k) > 0$ for all $j \in J$, while from Eq. (2.4), we have $f(x^k) < f(x^*)$ for k sufficiently large (the case where $x^k = x^*$ for infinitely many k is excluded by the assumption that $I \cup J$ is nonempty). Furthermore, the conditions $|h_i(x^k)| \le \epsilon m(x_k)$ for $i \notin I$ and $g_j(x^k) \le \epsilon m(x_k)$ for $j \notin J$ are equivalent to

$$|\lambda_i^k| \le \epsilon \min\bigl\{\min\{|\lambda_i^k| \mid i \in I\}, \min\{\mu_j^k \mid j \in J\}\bigr\}, \qquad \forall\, i \notin I,$$

and

$$\mu_j^k \le \epsilon \min\bigl\{\min\{|\lambda_i^k| \mid i \in I\}, \min\{\mu_j^k \mid j \in J\}\bigr\}, \qquad \forall\, j \notin J,$$

respectively, so they evidently hold for all sufficiently large k in $\mathcal{K}$. Since every neighborhood of x^* must contain all x^k with sufficiently large k in $\mathcal{K}$, this proves condition (iv). **Q.E.D.**

Note that if X is regular at x^*, i.e., $N_X(x^*) = T_X(x^*)^\perp$, condition (i) of Prop. 1 becomes $-\Bigl(\mu_0^* \nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla h_i(x^*) + \sum_{j=1}^r \mu_j^* \nabla g_j(x^*)\Bigr) \in T_X(x^*)^\perp$ or equivalently

$$\left(\mu_0^* \nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla h_i(x^*) + \sum_{j=1}^r \mu_j^* \nabla g_j(x^*)\right)' y \ge 0, \qquad \forall\, y \in T_X(x^*).$$

If in addition, the scalar μ_0^* can be shown to be strictly positive, then by normalization we can choose $\mu_0^* = 1$, and condition (i) of Prop. 1 becomes equivalent to the Lagrangian stationarity condition (1.4). Thus, if X is regular at x^* and we can guarantee that $\mu_0^* = 1$, the vector $(\lambda^*, \mu^*) = \{\lambda_1^*, \ldots, \lambda_m^*, \mu_1^*, \ldots, \mu_r^*\}$ is a Lagrange multiplier vector that satisfies condition (iv) of Prop. 1. A key fact is that this condition is stronger than the complementary slackness condition (1.6) [if $\mu_j^* > 0$, then according to condition (iv), the corresponding jth inequality constraint must be violated arbitrarily close to x^* [cf. Eq. (2.1)], implying that $g_j(x^*) = 0$]. The strengthening of the complementary slackness condition will turn out to be of crucial significance in the next section.

Condition (iv) of Prop. 1 also has another important implication. It asserts that the multiplier vector (λ^*, μ^*) has a special sensitivity-like property: through the signs of the λ_i^* and μ_j^*, it identifies the constraints that if violated, permit the reduction of the cost function. We call such a Lagrange multiplier *informative*, and in a separate, more extensive report on the subject of this paper, we show an interesting new result: *if $T_X(x^*)$ is convex and there exists at least one Lagrange multiplier vector at a local minimum x^* of problem (1.1)-(1.2), then there exists at least one Lagrange multiplier vector that is informative* [convexity of $T_X(x^*)$, which is guaranteed if X is regular at x^*, is an essential property for this result to hold].

To place Prop. 1 in perspective, we note that its line of proof, based on the quadratic penalty function, is due to McShane [McS73]. Hestenes [Hes75] observed that McShane's proof can be used to strengthen the complementary slackness condition to assert the existence, within any neighborhood B of x^*, of an $x \in B \cap X$ such that

$$\lambda_i^* h_i(x) > 0, \quad \forall\, i \in I, \qquad g_j(x) > 0, \quad \forall\, j \in J, \tag{2.11}$$

which is slightly weaker than condition (iv) of Prop. 1 [there is no requirement that x simultaneously satisfies $f(x) < f(x^*)$ and Eq. (2.2)]. McShane and Hestenes considered only the case where $X = \Re^n$. The case where X is a closed convex set was considered in Bertsekas [Ber99], where a generalized version of the Mangasarian-Fromovitz constraint qualification was also proved. The extension to the case where X is a general closed set and the strengthened version of condition (iv) are presented in the present paper for the first time.

3. PSEUDONORMALITY AND CONSTRAINT QUALIFICATIONS

Proposition 1 leads to the introduction of a general constraint qualification under which the scalar μ_0^* in Prop. 1 cannot be zero. In particular, let us say that a feasible vector x^* of problem (1.1)-(1.2) is *pseudonormal* if there are no scalars $\lambda_1, \ldots, \lambda_m, \mu_1, \ldots, \mu_r$, such that:

(i) $-\left(\sum_{i=1}^m \lambda_i \nabla h_i(x^*) + \sum_{j=1}^r \mu_j \nabla g_j(x^*)\right) \in N_X(x^*)$.

(ii) $\mu_j \geq 0$, for all $j = 1, \ldots, r$, and $\mu_j = 0$ for all $j \notin A(x^*)$.

(iii) In every neighborhood B of x^* there is an $x \in B \cap X$ such that

$$\sum_{i=1}^m \lambda_i h_i(x) + \sum_{j=1}^r \mu_j g_j(x) > 0. \tag{3.1}$$

It can be seen that if x^* is a pseudonormal local minimum, the Fritz John conditions of Prop. 1 cannot be satisfied with $\mu_0^* = 0$, so that μ_0^* can be taken equal to 1. Then, if X is regular at x^*, the vector $(\lambda^*, \mu^*) = (\lambda_1^*, \ldots, \lambda_m^*, \mu_1^*, \ldots, \mu_r^*)$ is an informative Lagrange multiplier.

Let us also mention another interesting property of the constraint set, called *quasinormality*, which is based on the condition (2.11). Quasinormality was introduced, for the case $X = \Re^n$, by Hestenes [Hes75], who showed how it can be used to unify various constraint qualifications. Pseudonormality implies quasinormality, and is thus intermediate between the constraint qualifications of this paper and quasinormality. As we will show in Section 4, pseudonormality is more naturally suited for showing results relating to the existence of exact penalties. Let us also note that in a separate report we generalize the notion of quasinormality to the case where $X \neq \Re^n$, and we discuss its relation to a slightly weaker version of pseudonormality where the vector x in Eq. (3.1) is additionally required to satisfy $\lambda_i h_i(x) \geq 0$ for all i and $\mu_j g_j(x) \geq 0$ for all j.

We now give some additional constraint qualifications, which together with CQ1-CQ3, given in Section 1, will be seen to imply pseudonormality of a feasible vector x^*.

CQ4: $X = \Re^n$ and for some integer $\overline{r} < r$, the following superset $\overline{C}$ of the constraint set C,

$$\overline{C} = \{x \mid h_i(x) = 0,\ i = 1, \ldots, m,\ g_j(x) \leq 0,\ j = \overline{r}+1, \ldots, r\},$$

is pseudonormal at x^*. Furthermore, there exists a $y \in \Re^n$ such that

$$\nabla h_i(x^*)' y = 0, \quad i = 1, \ldots, m, \qquad \nabla g_j(x^*)' y \leq 0, \quad \forall\ j \in A(x^*),$$

$$\nabla g_j(x^*)'y < 0, \quad \forall\ j \in \{1, \ldots, \overline{r}\} \cap A(x^*).$$

Since CQ1-CQ3 imply pseudonormality, a fact to be shown in the subsequent Prop. 2, we see that CQ4 generalizes all the constraint qualifications CQ1-CQ3.

CQ5:

(a) The equality constraints with index above some $\overline{m} \leq m$:

$$h_i(x) = 0, \qquad i = \overline{m} + 1, \ldots, m,$$

are linear.

(b) There does not exist a vector $\lambda = (\lambda_1, \ldots, \lambda_m)$ such that

$$\sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) \in N_X(x^*) \tag{3.2}$$

and at least one of the scalars $\lambda_1, \ldots, \lambda_{\overline{m}}$ is nonzero.

(c) The subspace

$$V_L(x^*) = \{y \mid \nabla h_i(x^*)'y = 0,\ i = \overline{m} + 1, \ldots, m\}$$

has a nonempty intersection with the interior of $N_X(x^*)^{\perp}$.

(d) There exists a $y \in N_X(x^*)^{\perp}$ such that

$$\nabla h_i(x^*)'y = 0, \quad i = 1, \ldots, m, \qquad \nabla g_j(x^*)'y < 0, \quad \forall\ j \in A(x^*).$$

We refer to CQ5 as the *generalized Mangasarian-Fromovitz constraint qualification*, since it reduces to CQ2 when $X = \Re^n$ and none of the equality constraints is assumed to be linear. This constraint qualification has several special cases, which we list below.

CQ5a:

(a) There does not exist a nonzero vector $\lambda = (\lambda_1, \ldots, \lambda_m)$ such that

$$\sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) \in N_X(x^*).$$

(b) There exists a $y \in N_X(x^*)^{\perp}$ such that

$$\nabla h_i(x^*)'y = 0, \quad i = 1, \ldots, m, \qquad \nabla g_j(x^*)'y < 0, \quad \forall\ j \in A(x^*).$$

CQ5b: There are no inequality constraints, the gradients $\nabla h_i(x^*)$, $i = 1, \ldots, m$, are linearly independent, and the subspace

$$V(x^*) = \{y \mid \nabla h_i(x^*)'y = 0,\ i = 1, \ldots, m\}$$

contains a point in the interior of $N_X(x^*)^\perp$.

CQ5c: X is convex, there are no inequality constraints, the functions h_i, $i = 1, \ldots, m$, are linear, and the linear manifold

$$\{y \mid h_i(x) = 0,\ i = 1, \ldots, m\}$$

contains a point in the interior of X.

CQ5d: X is convex, the functions g_j are convex, there are no equality constraints, and there exists a feasible vector $\overline{x}$ satisfying

$$g_j(\overline{x}) < 0, \qquad \forall\ j \in A(x^*).$$

CQ5a is the special case of CQ5 where all equality constraints are assumed nonlinear. CQ5b is a special case of CQ5 (where there are no inequality constraints and no linear equality constraints) based on the fact that if $\nabla h_i(x^*)$, $i = 1, \ldots, m$, are linearly independent and the subspace $V(x^*)$ contains a point in the interior of $N_X(x^*)^\perp$, then it can be shown that assumption (b) of CQ5 is satisfied. Finally, the convexity assumptions in CQ5c and CQ5d can be used to establish the corresponding assumption (c) and (d) of CQ5, respectively. Note that CQ5d is the well-known Slater constraint qualification, introduced in [Sla50].

Let us also mention the following constraint qualification.

CQ6: There are no scalars $\lambda_1, \ldots, \lambda_m, \mu_1, \ldots, \mu_r$ such that conditions (i) and (ii) of the definition of pseudonormality are satisfied.

CQ6 is the constraint qualification introduced by Rockafellar [Roc93], who used McShane's line of proof to derive the Fritz John conditions in the classical form where complementary slackness replaces condition (iv) in Prop. 1. Clearly CQ6 is a more restrictive condition than pseudonormality, since it does not require condition (iii) of the definition of pseudonormality. It can be shown that CQ6 together with regularity of X at x^*, is equivalent to CQ5a. This is proved by Rockafellar and Wets [RoW98] in the case where $X = \Re^n$, and can be verified in the more general case where $X \neq \Re^n$ by using their analysis given in p. 226 of [RoW98]. However, CQ3, CQ4, and CQ5 do not imply CQ6. Thus CQ6 is not as effective in unifying various existing constraint qualifications as pseudonormality, which is implied by all the constraint qualifications CQ1-CQ6, as shown in the following proposition.

Proposition 2: A feasible point x^* of problem (1.1)-(1.2) is pseudonormal if any one of the constraint qualifications CQ1-CQ6 is satisfied.

Proof: We will not consider CQ2 since it is a special case of CQ5. It is also evident that CQ6 implies pseudonormality. Thus we will prove the result for the cases CQ1, CQ3, CQ4, and CQ5 in that order. In all cases, the method of proof is by contradiction, i.e., we assume that there are scalars λ_i, $i = 1, \ldots, m$, and μ_j, $j = 1, \ldots, r$, which satisfy conditions (i)-(iii) of the definition of pseudonormality. We then assume that each of the constraint qualifications CQ1, CQ3, CQ4, and CQ5 is in turn also satisfied, and we arrive at a contradiction.

CQ1: Since $X = \Re^n$, implying that $N_X(x^*) = \{0\}$, and we have $\mu_j = 0$ for all $j \notin A(x^*)$ by condition (ii), we can write condition (i) as

$$\sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) + \sum_{j \in A(x^*)} \mu_j \nabla g_j(x^*) = 0.$$

Linear independence of $\nabla h_i(x^*)$, $i = 1, \ldots, m$, and $\nabla g_j(x^*)$, $j \in A(x^*)$, implies that $\lambda_i = 0$ for all i and $\mu_j = 0$ for all $j \in A(x^*)$. This, together with the condition $\mu_j = 0$ for all $j \notin A(x^*)$, violates condition (iii).

CQ3: By the linearity of h_i and the concavity of g_j, we have for all $x \in \Re^n$,

$$h_i(x) = h_i(x^*) + \nabla h_i(x^*)'(x - x^*), \qquad i = 1, \ldots, m,$$

$$g_j(x) \le g_j(x^*) + \nabla g_j(x^*)'(x - x^*), \qquad j = 1, \ldots, r.$$

By multiplying these two relations with λ_i and μ_j, and by adding over i and j, respectively, we obtain

$$\begin{aligned} \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{r} \mu_j g_j(x) &\le \sum_{i=1}^{m} \lambda_i h_i(x^*) + \sum_{j=1}^{r} \mu_j g_j(x^*) \\ &\quad + \left(\sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) + \sum_{j=1}^{r} \mu_j \nabla g_j(x^*) \right)' (x - x^*) \qquad (3.3) \\ &= 0, \end{aligned}$$

where the last equality holds because we have $\lambda_i h_i(x^*) = 0$ for all i and $\mu_j g_j(x^*) = 0$ for all j [by condition (ii)], and

$$\sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) + \sum_{j=1}^{r} \mu_j \nabla g_j(x^*) = 0$$

[by condition (i)]. On the other hand, by condition (iii), there is an x satisfying $\sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{r} \mu_j g_j(x) > 0$, contradicting Eq. (3.3).

CQ4: It is not possible that $\mu_j = 0$ for all $j \in \{1, \ldots, \overline{r}\}$, since if this were so, the pseudonormality assumption for $\overline{C}$ would be violated. Thus we have $\mu_j > 0$ for some $j \in \{1, \ldots, \overline{r}\} \cap A(x^*)$. It follows that for the vector y appearing in the statement of CQ4, we have $\sum_{j=1}^{\overline{r}} \mu_j \nabla g_j(x^*)'y < 0$, so that

$$\sum_{i=1}^{m} \lambda_i \nabla h_i(x^*)'y + \sum_{j=1}^{r} \mu_j \nabla g_j(x^*)'y < 0.$$

This contradicts the equation

$$\sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) + \sum_{j=1}^{r} \mu_j \nabla g_j(x^*) = 0,$$

[cf. condition (i)].

CQ5: We first show by contradiction that at least one of the $\lambda_1, \ldots, \lambda_{\overline{m}}$ and μ_j, $j \in A(x^*)$ must be nonzero. If this were not so, then by using a translation argument we may assume that x^* is the origin, and the linear constraints have the form $a_i'x = 0$, $i = \overline{m}+1, \ldots, m$. Using condition (i) we have

$$-\sum_{i=\overline{m}+1}^{m} \lambda_i a_i \in N_X(x^*). \tag{3.4}$$

Let $\overline{y}$ be the interior point of $N_X(x^*)^\perp$ that satisfies $a_i'\overline{y} = 0$ for all $i = \overline{m}+1, \ldots, m$, and let S be an open sphere centered at the origin such that $\overline{y} + d \in N_X(x^*)^\perp$ for all $d \in S$. We have from Eq. (3.4),

$$\sum_{i=\overline{m}+1}^{m} \lambda_i a_i' d \geq 0, \qquad \forall\, d \in S,$$

from which we obtain $\sum_{i=\overline{m}+1}^{m} \lambda_i a_i = 0$. This contradicts condition (iii), which requires that there exists some $x \in S \cap X$ such that $\sum_{i=\overline{m}+1}^{m} \lambda_i a_i' x > 0$.

Next we show by contradiction that we cannot have $\mu_j = 0$ for all j. If this were so, by condition (i) there must exist a nonzero vector $\lambda = (\lambda_1, \ldots, \lambda_m)$ such that

$$-\sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) \in N_X(x^*). \tag{3.5}$$

By what has been proved above, the multipliers $\lambda_1, \ldots, \lambda_{\overline{m}}$ of the nonlinear constraints cannot be all zero, so Eq. (3.5) contradicts assumption (b) of CQ5.

Hence we must have $\mu_j > 0$ for at least one j, and since $\mu_j \geq 0$ for all j with $\mu_j = 0$ for $j \notin A(x^*)$, we obtain

$$\sum_{i=1}^{m} \lambda_i \nabla h_i(x^*)' y + \sum_{j=1}^{r} \mu_j \nabla g_j(x^*)' y < 0,$$

for the vector y of $N_X(x^*)^\perp$ that appears in assumption (d) of CQ5. Thus,

$$-\left(\sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) + \sum_{j=1}^{r} \mu_j \nabla g_j(x^*)\right) \notin \left(N_X(x^*)^\perp\right)^\perp.$$

Since $N_X(x^*) \subset \left(N_X(x^*)^\perp\right)^\perp$, this contradicts condition (i). **Q.E.D.**

A consequence of Prop. 2 is that each of the constraint qualifications CQ1-CQ6 implies that x^* is pseudonormal, so if X is regular at x^*, by Prop. 1, the constraint set C admits informative Lagrange multipliers at x^*. In the next two sections, we will also show similar implications regarding the admittance of an exact penalty at x^*.

4. PSEUDONORMALITY AND EXISTENCE OF EXACT PENALTY FUNCTIONS

We will show that pseudonormality implies that the constraint admits an exact penalty, which in turn, together with regularity of X at x^*, implies that the constraint set admits Lagrange multipliers. We first use the generalized Mangasarian-Fromovitz constraint qualification CQ5 to obtain a necessary condition for a local minimum of the exact penalty function.

Proposition 3: Let x^* be a local minimum of

$$F_c(x) = f(x) + c\left(\sum_{i=1}^{m} |h_i(x)| + \sum_{j=1}^{r} g_j^+(x)\right)$$

over X. Then there exist $\lambda_1^*, \ldots, \lambda_m^*$ and $\mu_1^*, \ldots, \mu_r^*$ such that

$$-\left(\nabla f(x^*) + c\left(\sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*) + \sum_{j=1}^{r} \mu_j^* \nabla g_j(x^*)\right)\right) \in N_X(x^*),$$

$$\lambda_i^* = 1 \ \text{ if } h_i(x^*) > 0, \qquad \lambda_i^* = -1 \ \text{ if } h_i(x^*) < 0, \qquad \lambda_i^* \in [-1,1] \ \text{ if } h_i(x^*) = 0,$$

$$\mu_j^* = 1 \ \text{ if } g_j(x^*) > 0, \qquad \mu_j^* = 0 \ \text{ if } g_j(x^*) < 0, \qquad \mu_j^* \in [0,1] \ \text{ if } g_j(x^*) = 0.$$

Proof: The problem of minimizing $F_c(x)$ over $x \in X$ can be converted to the problem

$$\text{minimize } \ f(x) + c\left(\sum_{i=1}^{m} w_i + \sum_{j=1}^{r} v_j\right)$$

$$\text{subject to } \ x \in X, \ \ h_i(x) \le w_i, \ -h_i(x) \le w_i, \ i = 1, \ldots, m, \quad g_j(x) \le v_j, \ 0 \le v_j, \ j = 1, \ldots, r,$$

which involves the auxiliary variables w_i and v_j. It can be seen that at the local minimum of this problem that corresponds to x^* the constraint qualification CQ5 is satisfied. Thus, by Prop. 2, there exist multipliers satisfying the conditions of Prop. 1, which with straightforward calculation, yield scalars $\lambda_1^*, \ldots, \lambda_m^*$ and $\mu_1^*, \ldots, \mu_r^*$, satisfying the desired conditions. **Q.E.D.**

Proposition 4: If x^* is a feasible vector of problem (1.1)-(1.2), which is pseudonormal, the constraint set admits an exact penalty at x^*.

Proof: Assume the contrary, i.e., that there exists a continuously differentiable f such that x^* is a strict local minimum of f over the constraint set C, while x^* is not a local minimum over $x \in X$ of the function

$$F_k(x) = f(x) + k\left(\sum_{i=1}^{m} |h_i(x)| + \sum_{j=1}^{r} g_j^+(x)\right)$$

for all $k = 1, 2, \ldots$ Let $\epsilon > 0$ be such that

$$f(x^*) < f(x), \qquad \forall\, x \in C \text{ with } x \ne x^* \text{ and } \|x - x^*\| \le \epsilon. \tag{4.1}$$

Suppose that x^k minimizes $F_k(x)$ over the (compact) set of all $x \in X$ satisfying $\|x - x^*\| \le \epsilon$. Then, since x^* is not a local minimum of F_k over X, we must have that $x^k \ne x^*$, and that x^k is infeasible for problem (1.2), i.e.,

$$\sum_{i=1}^{m} |h_i(x^k)| + \sum_{j=1}^{r} g_j^+(x^k) > 0. \tag{4.2}$$

We have

$$F_k(x^k) = f(x^k) + k\left(\sum_{i=1}^{m} |h_i(x^k)| + \sum_{j=1}^{r} g_j^+(x^k)\right) \le f(x^*), \tag{4.3}$$

so it follows that $h_i(x^k) \to 0$ for all i and $g_j^+(x^k) \to 0$ for all j. The sequence $\{x^k\}$ is bounded and if $\overline{x}$ is any of its limit points, we have that $\overline{x}$ is feasible. From Eqs. (4.1) and (4.3) it then follows that $\overline{x} = x^*$. Thus $\{x^k\}$ converges to x^* and we have $\|x^k - x^*\| < \epsilon$ for all sufficiently large k. This implies the following necessary condition for optimality of x^k (cf. Prop. 3):

$$-\left(\frac{1}{k}\nabla f(x^k) + \sum_{i=1}^{m} \lambda_i^k \nabla h_i(x^k) + \sum_{j=1}^{r} \mu_j^k \nabla g_j(x^k)\right) \in N_X(x^k), \tag{4.4}$$

where

$$\lambda_i^k = 1 \quad \text{if } h_i(x^k) > 0, \qquad \lambda_i^k = -1 \quad \text{if } h_i(x^k) < 0, \qquad \lambda_i^k \in [-1,1] \quad \text{if } h_i(x^k) = 0,$$

$$\mu_j^k = 1 \quad \text{if } g_j(x^k) > 0, \qquad \mu_j^k = 0 \quad \text{if } g_j(x^k) < 0, \qquad \mu_j^k \in [0,1] \quad \text{if } g_j(x^k) = 0.$$

In view of Eq. (4.2), we can find a subsequence $\{\lambda^k, \mu^k\}_{k\in\mathcal{K}}$ such that for some equality constraint index i we have $|\lambda_i^k| = 1$ and $h_i(x^k) \neq 0$ for all $k \in \mathcal{K}$ or for some inequality constraint index j we have $\mu_j^k = 1$ and $g_j(x^k) > 0$ for all $k \in \mathcal{K}$. Let (λ, μ) be a limit point of this subsequence. We then have $(\lambda, \mu) \neq (0, 0)$, $\mu \ge 0$. Using the closure of the mapping $x \mapsto N_X(x)$, Eq. (4.4) yields

$$-\left(\sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) + \sum_{j=1}^{r} \mu_j \nabla g_j(x^*)\right) \in N_X(x^*). \tag{4.5}$$

Finally, for all $k \in \mathcal{K}$, we have $\lambda_i^k h_i(x^k) \ge 0$ for all i, $\mu_j^k g_j(x^k) \ge 0$ for all j, so that, for all $k \in \mathcal{K}$, $\lambda_i h_i(x^k) \ge 0$ for all i, $\mu_j g_j(x^k) \ge 0$ for all j. Since by construction of the subsequence $\{\lambda^k, \mu^k\}_{k\in\mathcal{K}}$, we have for some i and all $k \in \mathcal{K}$, $|\lambda_i^k| = 1$ and $h_i(x^k) \neq 0$, or for some j and all $k \in \mathcal{K}$, $\mu_j^k = 1$ and $g_j(x^k) > 0$, it follows that for all $k \in \mathcal{K}$,

$$\sum_{i=1}^{m} \lambda_i h_i(x^k) + \sum_{j=1}^{r} \mu_j g_j(x^k) > 0. \tag{4.6}$$

Thus, Eqs. (4.5) and (4.6) violate the hypothesis that x^* is pseudonormal. **Q.E.D.**

Proposition 5: Let x^* be a feasible vector of problem (1.1)-(1.2), and let X be regular at x^*. If the constraint set admits an exact penalty at x^*, it admits Lagrange multipliers at x^*.

Proof: Suppose that a given continuously differentiable function $f(x)$ has a local minimum at x^*. Then the function $f(x) + \|x - x^*\|^2$ has a strict local minimum at x^*. Since C admits an exact penalty at x^*, there exist λ_i^* and μ_j^* satisfying the conditions of Prop. 3 (the term $\|x - x^*\|^2$ in the cost function is inconsequential, since its gradient at x^* is 0). In view of the regularity of X at x^*, the λ_i^* and μ_j^* are Lagrange multipliers. **Q.E.D.**

REFERENCES

[AuF90] Aubin, J.-P., and Frankowska, H., Set-Valued Analysis, Birkhauser, Boston.

[BSS93] Bazaraa, M. S., Sherali, H. D., and Shetty, C. M., 1993. Nonlinear Programming Theory and Algorithms, (2nd Ed.), Wiley, N. Y.

[BaG82] Bazaraa, M. S., and Goode, J. J., 1982. "Sufficient Conditions for a Globally Exact Penalty Function without Convexity," Math. Programming Stud., Vol. 19, pp. 1-15.

[Ber99] Bertsekas, D. P., "Nonlinear Programming: 2nd Edition," Athena Scientific, Belmont, MA, 1999.

[BoL00] Borwein, J. M., and Lewis, A. S., Convex Analysis and Nonlinear Optimization, Springer-Verlag, N. Y.

[BoS00] Bonnans, J. F., and Shapiro, A., 2000. Perturbation Analysis of Optimization Problems, Springer-Verlag, N. Y.

[HaM79] Han, S. P., and Mangasarian, O. L., 1979. "Exact Penalty Functions in Nonlinear Programming," Math. Programming, Vol. 17, pp. 251-269.

[Hes75] Hestenes, M. R., 1975. Optimization Theory: The Finite Dimensional Case, Wiley, N. Y.

[Joh48] John, F., 1948. "Extremum Problems with Inequalities as Subsidiary Conditions," in Studies and Essays: Courant Anniversary Volume, K. O. Friedrichs, Neugebauer, O. E., and Stoker, J. J., (Eds.), Wiley-Interscience, N. Y., pp. 187-204.

[McS73] McShane, E. J., 1973. "The Lagrange Multiplier Rule," Amer. Mathematical Monthly, Vol. 80, pp. 922-925.

[Mor76] Mordukhovich, B. S., 1976. "Maximum Principle in the Problem of Time Optimal Response with Nonsmooth Constraints," J. of Applied Mathematics and Mechanics, Vol. 40, pp. 960-969.

[Pie69] Pietrzykowski, T., 1969. "An Exact Potential Method for Constrained Maxima," SIAM J. Numer. Anal., Vol. 6, pp. 294-304.

[RoW98] Rockafellar, R. T., and Wets, R. J.-B., 1998. Variational Analysis, Springer-Verlag, Berlin.

[Roc93] Rockafellar, R. T., 1993. "Lagrange Multipliers and Optimality," SIAM Review, Vol. 35, pp. 183-238.

[Zan67] Zangwill, W. I., 1967. "Nonlinear Programming via Penalty Functions," Management Sci., Vol. 13, pp. 344-358.

Slowest Descent Sequence Estimation for Problems with a Quadratic Log-Likelihood *

Predrag Spasojević and Costas N. Georghiades†
WINLAB, ECE Department, Rutgers University
Piscataway, NJ 08854-8060
spasojev@winlab.rutgers.edu

Abstract

A new approach to sequence estimation is proposed and its performance is analyzed for several of channels of practical interest. The proposed approach, termed the slowest descent method, is closely related to the zero-forcing equalizer for intersymbol interference (ISI) channels and to the decorrelator for the multi-user detection (MUD) problem. The latter two methods quantize the *unconstrained* sequence that maximizes the likelihood function. The proposed method refines the decorrelator and the zero-forcing estimates by comparing their likelihood to a set of discrete sequences along mutually orthogonal lines of the least decrease in the likelihood function. The slowest descent method is shown by simulations that it can provide a performance comparable to the performance of the maximum-likelihood estimator for the ISI channel and the multi-user detection problem. Both of these problems can be described using a quadratic log-likelihood function.

1 Introduction

It is well known that the problem of optimum sequence estimation is in general computationally intractable. Exceptions are problems whose structure allows the use of the Viterbi algorithm, and, of course, the case of symbol-by-symbol detection for memoryless channels without coding. Much effort has been devoted over the years to the problem of suboptimum techniques that produce a performance close to optimum.

The proposed approach, termed the slowest descent method, is closely related to the zero-forcing equalizer for intersymbol interference (ISI) channels and to the decorrelator for the multi-user detection (MUD) problem. The latter two methods quantize the *unconstrained* sequence that maximizes the likelihood function. The proposed method refines the decorrelator and the zero-forcing estimates by comparing their likelihood to a set of discrete sequences along mutually orthogonal lines of the least decrease in the likelihood function and selecting the best. We refer to this search technique as the slowest descent method. The method suggests analyzing the slope of the continuous log-likelihood function in a neighborhood of the unconstrained maximizer to determine the slowest descent

---

*This work was supported in part by National Science Foundation grant NCR-9314221.

†With the Electrical Engineering Department, Texas A&M University, College Station, TX 77843-3128. E-mail: georghia@ee.tamu.edu.

lines. The discrete sequences along these lines are easily found in the case of binary antipodal signaling and form a discrete sequence set of size linear in the sequence length.

Section 2 introduces the slowest descent method. Section 3 describes the general channel model, the quadratic log-likelihood function for this model. and sequence estimation using the slowest descent method. In Sections 4 and 5 the proposed approach is applied to two example channels: the multi-access interference, and the ISI channel. Section 6 concludes.

2 The Slowest Descent Method

For problems for which a log-likelihood function $l(\mathbf{a})$ of an M-dimensional discrete valued sequence can be formed, the optimal estimator selects the following (maximum-likelihood) sequence:

$$\hat{\mathbf{a}}_d = \arg\max_{\mathbf{a}\in\mathbf{\Omega}_d} l(\mathbf{a}), \tag{1}$$

where $\mathbf{\Omega}_d$ is the discrete set of admissible sequences. Sequences will be represented as Mx1 vectors $\mathbf{a} = [a_1, \ldots, a_M]^T$ in the real space $\mathcal{R}^M$ M-dimensional vector space.

In this paper we assume that $l(\mathbf{a})$ is well defined over $\mathcal{R}^M \supset \mathbf{\Omega}_d$, and an optimal solution of the "continuous" optimization problem

$$\hat{\mathbf{a}}_c = \arg\max_{\mathbf{a}\in\mathcal{R}^M} l(\mathbf{a}) \tag{2}$$

exists. It is also assumed that the Hessian of $l(\mathbf{a})$, $\mathbf{H}^l(\mathbf{a}) \triangleq \frac{\partial^2 l(\mathbf{a})}{\partial\mathbf{a}\partial\mathbf{a}^T}$, is well defined at the continuous maximizer sequence $\hat{\mathbf{a}}_c$. Note that if $\hat{\mathbf{a}}_c$ is the local unconstrained maximizer of $l(\mathbf{a})$ then $\mathbf{H}^l(\hat{\mathbf{a}}_c)$ is a non-positive definite matrix.

A locally concave function $l(\mathbf{a})$ has a quadratic approximation in some neighborhood $U(\hat{\mathbf{a}}_c)$ as follows:

$$l(\mathbf{a}) \;\approx\; l(\hat{\mathbf{a}}_c) + \frac{1}{2}(\mathbf{a}-\hat{\mathbf{a}}_c)^T\mathbf{H}^l(\hat{\mathbf{a}}_c)(\mathbf{a}-\hat{\mathbf{a}}_c). \tag{3}$$

It is well known (see, e.g., [1]) that, within the neighborhood $U(\hat{\mathbf{a}}_c)$, the eigenvectors $\hat{\mathbf{\Delta}}^i_{min}$ of $\mathbf{H}^l(\hat{\mathbf{a}}_c)$ are mutually orthogonal solutions of the following problem:

$$\hat{\mathbf{\Delta}}^i_{min} = \arg\min_{\mathbf{\Delta}} \{l(\hat{\mathbf{a}}_c) - l(\hat{\mathbf{a}}_c + \epsilon\mathbf{\Delta})\} \tag{4}$$

for $i = 1, \ldots, M$. Here $\mathbf{\Delta}$ is a unit-length vector and the constant $\epsilon > 0$ is such that $\hat{\mathbf{a}}_c + \epsilon\mathbf{\Delta} \in U(\hat{\mathbf{a}}_c)$ for any unit-length $\mathbf{\Delta}$. That is, the eigenvectors $\{\hat{\mathbf{\Delta}}^i_{min}\}$ define mutually orthogonal directions of the least decrease in $l(\mathbf{a})$ from the maximizer $\hat{\mathbf{a}}_c$ in $U(\hat{\mathbf{a}}_c)$. $\hat{\mathbf{\Delta}}^1_{min}$ defines the least decrease direction, $\hat{\mathbf{\Delta}}^2_{min}$ defines the second-least decrease direction, etc. Since $l(\hat{\mathbf{a}}_c + \epsilon\mathbf{\Delta}) = l(\hat{\mathbf{a}}_c - \epsilon\mathbf{\Delta})$ for a sufficiently small ϵ, we can talk about M mutually orthogonal straight lines of the least decrease of $l(\mathbf{a})$ defined by $\hat{\mathbf{a}}_c$ $\{\hat{\mathbf{\Delta}}^i_{min}\}$ as:

$$\{\hat{\mathbf{a}}_c + \rho\hat{\mathbf{\Delta}}^i_{min} : \rho \in \mathcal{R}, i = 1, \ldots, M\}. \tag{5}$$

The eigenvector which defines the line of the least decrease in $l(\mathbf{a})$ from the maximizer corresponds to the smallest absolute eigen-value of $\mathbf{H}^l(\hat{\mathbf{a}}_c)$, is termed the smallest eigenvector, and is denoted with $\hat{\mathbf{\Delta}}_{min} \triangleq \hat{\mathbf{\Delta}}^1_{min}$. Its corresponding line in $\mathcal{R}^M$ is termed the slowest descent line. In the same manner, the eigenvectors that correspond to P smallest

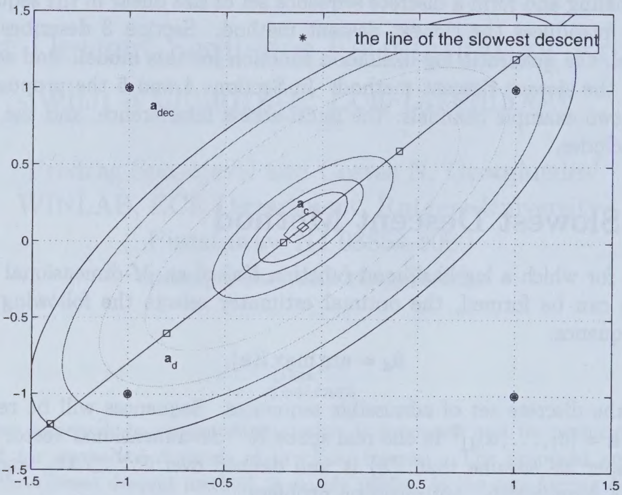


Figure 1: Contour plot and line of slowest descent for $M = 2$.

absolute eigen-values are termed P smallest eigenvectors and their corresponding lines are termed P (mutually orthogonal) lines of the slowest descent.

Figure 1 depicts the case of a quadratic objective function for a $M = 2$ sequence set. The ellipsoids of equal function value in the two-dimensional sequence space are elongated the most in the direction of the least decrease in the objective function, i.e., in the direction determined by the smallest eigenvector of the Hessian. They are centered at $\hat{\mathbf{a}}_c$ and their axis of largest elongation is the slowest descent line. $\mathbf{\Omega}_d = \{[-1,-1]^T, [-1,1]^T, [1,-1]^T, [1,1]^T\}$ and $\hat{\mathbf{a}}_d = [-1,-1]^T$ in this case. The slowest descent line passes in the neighborhood of the two most likely sequences $\hat{\mathbf{a}}_d$ and $[1,1]^T$. By comparing the value of the function at $M+1 = 3$ discrete sequences closest to the slowest descent line we can determine $\hat{\mathbf{a}}_d$.

The slowest descent method chooses as the sequence estimate $\hat{\mathbf{a}}_d \in \mathbf{\Omega}_d$ the most likely of the sequences closest in the l_2 norm to $P \leq M$ smallest descent lines in $\mathcal{R}^M$. For arbitrary log-likelihood functions (not necessarily quadratic) the special case of the slowest descent approach for $P = 0$, that is, when the sequence estimate is the sequence closest to $\hat{\mathbf{a}}_c$, will be termed the generalized decorrelator method. Note that, in Figure 1, the generalized decorrelator method would (erroneously) choose the discrete sequence $\hat{\mathbf{a}}_{dec} = \text{sign}[\hat{\mathbf{a}}_c] = [-1,1]^T \neq \hat{\mathbf{a}}_d$, where the sign is taken element by element.

2.1 Binary Antipodal Sequences Closest to a Line in $\mathcal{R}^M$

Here we assume that BPSK is employed, in which case the components of $\mathbf{a}$ are in $\{-1,+1\}$. Generalization to other signal constellations is addressed briefly in [2].

In order to find all the sequences in $\mathbf{\Omega}_d$ closest to a line $(\hat{\mathbf{a}}_c + \varrho\mathbf{\Delta})$, where ϱ is a real number and $\boldsymbol{\Delta}$ is a vector in $\mathcal{R}^M$, all distinct sequences that satisfy

$$\mathbf{a}_d^{\varrho} = \text{sign}[\hat{\mathbf{a}}_c + \varrho\boldsymbol{\Delta}] \tag{6}$$

must be determined. It is easily observed that there can be at most $(M+1)$ distinct sequences $\mathbf{a}_d^{\varrho}$ satisfying this equality. They can be determined by those values of ϱ for which there is a change in the sign of any coordinate of $(\hat{\mathbf{a}}_c + \varrho\boldsymbol{\Delta})$. The values of ϱ for which the change of sign occurs are

$$\varrho^k = \hat{a}_{c,k}/\Delta_k \tag{7}$$

(assuming $\Delta_k \neq 0$) where $\hat{a}_{c,k}$ and Δ_k are k-th elements of the respective vectors $\hat{\mathbf{a}}_c$ and $\boldsymbol{\Delta}$. Note that $\text{Prob}\{\Delta_k = 0 \text{ or } \hat{a}_{c,m} = 0, \text{ for any } k \text{ or } m\} = 0$. Thus, to simplify notation we assume that $\Delta_k \neq 0$ and $\hat{a}_{c,k} \neq 0$ for all k.

Let's define a $\{1,\ldots,M\} \to \{1,\ldots,M\}$ mapping

$$m \to k(m): \quad \varrho^{k(m)} < \varrho^{k(m+1)}. \tag{8}$$

This mapping is determined by an increasing ordering of ρ^k in (7). m_c is defined as the index for which $\varrho^{k(m_c-1)} < 0 < \varrho^{k(m_c)}$ holds

Let $\mathbf{e}_k$ denote the unit-length vector whose coordinates are all zero except the k-th coordinate which is 1, $k_l \triangleq k(m_c + l)$, and

$$\mathbf{a}^0 \triangleq \text{sign}[\hat{\mathbf{a}}_c]. \tag{9}$$

Now, the $M+1$ binary sequences closest to the line $\hat{\mathbf{a}}_c + \varrho\boldsymbol{\Delta}$ are:

$$\mathbf{a}^{l+1} = \mathbf{a}^l - 2a_{k_l}^0 \mathbf{e}_{k_l} \tag{10}$$

for $l = 0,\ldots,M-m_c$, and

$$\mathbf{a}^l = \mathbf{a}^{l+1} - 2a_{k_l}^0 \mathbf{e}_{k_l} \tag{11}$$

for $l = -1,\ldots,1-m_c$, where a_k^0 is the k-th coordinate of the vector $\mathbf{a}^0$.

For a quadratic $l(\mathbf{a})$ (or its quadratic approximation) with a Hessian $\mathbf{H}$, the log-likelihoods of binary sequences $\{\mathbf{a}^l\}$ can be computed as follows:

$$l(\mathbf{a}^{l+1}) = l(\mathbf{a}^l) + 2a_{k_l}^0(\hat{\mathbf{a}}_c - \mathbf{a}^l)^T\mathbf{h}_{k_l} - 2h_{k_l k_l} \tag{12}$$

for $l = 0,\ldots,M-m_c$, and

$$l(\mathbf{a}^l) = l(\mathbf{a}^{l+1}) + 2a_{k_{l+1}}^0(\hat{\mathbf{a}}_c - \mathbf{a}^{l+1})^T\mathbf{h}_{k_{l+1}} - 2h_{k_{l+1}k_{l+1}} \tag{13}$$

for $l = -1,\ldots,1-m_c$. Here, $\mathbf{h}_{k_l} = \mathbf{H}\mathbf{e}_{k_l}$, and $h_{k_l k_l} = \mathbf{h}_{k_l}^T\mathbf{e}_{k_l}$.

Note that, in our application, only the relative likelihood of the sequences $\{\mathbf{a}^l\}$ needs to be known; thus, we can formally set $l(\mathbf{a}^0) = 0$.

2.2 The Unconstrained Maximizer and Directions of Slowest Descent

For quadratic log-likelihood functions, the unconstrained maximizer has a closed form representation which is linear in the sufficient statistic. The log-likelihood Hessian is, in this case, a constant symmetric matrix and $U(\hat{\mathbf{a}}_c) = \mathcal{R}^{\mathcal{M}}$. Furthermore, the eigenvectors of $\mathbf{H}^l(\hat{\mathbf{a}}_c)$ are not functions of the received signal. Both the linear operator required for estimation of $\hat{\mathbf{a}}_c$ and the eigenvectors of $\mathbf{H}^l(\hat{\mathbf{a}}_c)$ can be pre-computed.

The known ISI channel, the multi-user channel with known user sequences, and the channel with Gaussian interference of known statistics are examples of problems with a quadratic log-likelihood function. The application of the slowest descent method to the first two applications is studied in Sections 4 and 5. The non-quadratic log-likelihood functions have been treated in detail in [2].

2.3 Method Summary

In short, the slowest descent method requires determining the following:

- Unconstrained maximizer $\hat{\mathbf{a}}_c$ in (2).
- P lines of the slowest descent in (5).
- $PM+1$ sequences closest to the slowest descent lines using (10) and (11).
- The best discrete sequence from the selected subset using (12) and (13).

3 Signal Model and the Log-Likelihood Function

We assume K users each transmitting a sequence of N symbols. The system $NK\times 1$ pulse vector $\mathbf{h}(t) = [\mathbf{h}_1^T(t), \ldots, \mathbf{h}_K^T(t)]^T$ incorporates all user pulses, i.e., the m-th coordinate of vector $\mathbf{h}_k(t)$ is $h_{k,m}(t) = h_k(t-mT)$ is the pulse corresponding to the m-th symbol of User k. The symbol sequence vector has the same form, $\mathbf{a} = [\mathbf{a}_1^T, \ldots, \mathbf{a}_K^T]^T$, where the m-th coordinate of vector $\mathbf{a}_k$ is $a_{k,m}$, the m-th symbol of User k. The received signal is

$$r(t) = \sum_{k=1}^{K}\sum_{n=0}^{N-1} a_{k,n}h_k(t-nT) + n(t) = \mathbf{a}^T\mathbf{h}(t) + n(t), \tag{14}$$

where $n(t)$ is AWGN with a power spectral density $N_0/2$.

For this model the log-likelihood function (pre-multiplied by N_0) is

$$l(\mathbf{a}) = 2\,\mathbf{r}^T\mathbf{a} - \mathbf{a}^T\mathbf{Y}\mathbf{a}. \tag{15}$$

$$\mathbf{Y} =< \mathbf{h}(t), \mathbf{h}^T(t) > \stackrel{\triangle}{=} \int_{T_i}^{T_f} \mathbf{h}(t)\mathbf{h}^T(t)dt \tag{16}$$

is a KNxKN matrix and $[T_i, T_f]$ is the observation interval.

$$\mathbf{r} =< \mathbf{h}(t), r(t) > \tag{17}$$

is the sufficient statistic for the problem.

The optimal sequence is chosen using (1) where $\mathbf{\Omega}_d = \{1, -1\}^{KN}$. Clearly, in the most general case, optimal detection is a highly complex optimization problem. A solution is attempted using the slowest descent method.

The log-likelihood is quadratic as in (3), where $\mathbf{H}^l(\mathbf{a}) = -\mathbf{Y}$ and $\hat{\mathbf{a}}_c = [\mathbf{Y}]^{-1}\mathbf{r}$. For simulations described in Sections 4 and 5, the eigenvectors of $\mathbf{Y}$ are precomputed using its singular value decomposition. $\hat{\mathbf{a}}_c$ and the eigenvectors of $\mathbf{Y}$ define the mutually orthogonal lines of the slowest descent in $l(\mathbf{a})$. Sequence estimation, in this case is fully an open-loop method described in the previous section. Note that the computation of $\hat{\mathbf{a}}_c$ requires a matrix-vector multiplication which can be simplified for some special structures of $\mathbf{Y}$ and that eigenvectors of $\mathbf{Y}$ can be precomputed.

4 Multi-User Detection

For simplicity, synchronous direct sequence spread spectrum (DS-SS) signaling with known user sequences and equal user signal amplitudes, i.e., $E_b = E_{b,k} = \|\mathbf{h}_k(t)\|^2$ for all k, are assumed. Generalization to the asynchronous model with unequal amplitudes is straightforward. The user signal pulses can be described as

$$h_k(t) = \sum_{m=1}^{G} c_{m,k}\psi(t - m\frac{T}{G}) \triangleq \mathbf{s}_k^T \boldsymbol{\psi}(t), \tag{18}$$

where $\mathbf{s}_k \triangleq [c_{1,k}, \ldots, c_{G,k}]^T$ is the spreading sequence vector of User k, $\boldsymbol{\psi}(t) \triangleq [\psi(t), \psi(t-\frac{T}{G}), \ldots, \psi(t-(K-1)\frac{T}{G})]^T$ is the vector of chip pulses, and G is the spreading gain; pulses $\psi(t - m\frac{T}{G})$ are assumed to be orthonormal, that is, $< \boldsymbol{\psi}(t), \boldsymbol{\psi}^T(t) >= \mathbf{I}$.

For synchronous signaling it is sufficient to observe one symbol interval at a time; that is, the optimal receiver detects one $K-$dimensional symbol vector at a time.

For one observed symbol interval, i.e., $N = 1$, it is easy to show that $\mathbf{h}(t) = \mathbf{S}^T\boldsymbol{\psi}(t)$, where $\mathbf{S}$ is a $G \times K$ signature matrix whose k-th column is the signature vector $\mathbf{s}_k$ of User k. Furthermore, $\mathbf{Y} = \mathbf{S}^T\mathbf{S}$ and $\mathbf{r} = \mathbf{S}^T < \boldsymbol{\psi}(t), r(t) >$. That is, $\mathbf{r}$ is the sampled matched filter output. The symbol vector $\mathbf{a} \triangleq [a_1, \ldots, a_K]^T$, where the symbol interval index is suppressed, consists of user symbols for a given signaling interval. The Gram matrix of the signature sequence vectors $\mathbf{Y}$ is not banded in the general case. The unconstrained maximizer is, in this case, the non-quantized decorrelator estimate

$$\hat{\mathbf{a}}_c = \mathbf{Y}^{-1}\mathbf{r} = \left(\mathbf{S}^T\mathbf{S}\right)^{-1}\mathbf{S}^T < \boldsymbol{\psi}(t), r(t) > \quad \text{with} \quad \hat{\mathbf{a}}_{dec} = \text{sign}[\hat{\mathbf{a}}_c], \tag{19}$$

being the decorrelator estimate [3]. The eigenvectors of the Hessian

$$\mathbf{H}^l(\hat{\mathbf{a}}_c) = -\mathbf{Y} = \mathbf{S}^T\mathbf{S} \tag{20}$$

can be obtained from the singular value decomposition of $\mathbf{S}$. As stated in Section 2, the slowest descent method can not have a worse performance than the linear decorrelator detector, since it assures that the likelihood of the estimate is not smaller than the likelihood of $\hat{\mathbf{a}}_{dec}$.

A BPSK system with a processing gain $G = 7$ and $K = 6$ users of equal power is analyzed. The user of interest is assumed to be User 2.

Two examples of $\mathbf{S}$ are studied. Figure 2 demonstrates that it is sufficient to compare the likelihood of discrete symbol vectors $\mathbf{a} \in \Omega_d$ closest to the slowest likelihood descent line to obtain almost optimal performance for a randomly chosen (but fixed) signature matrix with the corresponding Gram matrix

$$\mathbf{Y} = \frac{E_b}{7}\begin{bmatrix} 7 & 3 & 1 & -1 & 1 & -1 \\ 3 & 7 & -3 & 3 & 1 & -1 \\ 1 & -3 & 7 & -3 & 3 & -3 \\ -1 & 3 & -3 & 7 & -3 & -1 \\ 1 & 1 & 3 & -3 & 7 & -3 \\ -1 & -1 & -3 & -1 & -3 & 7 \end{bmatrix}. \tag{21}$$

The parallel interference canceler (PIC) initialized with a decorrelator (PIC-DEC) has a significantly worse performance in this case. The bit error rate of the PIC-DEC oscillates

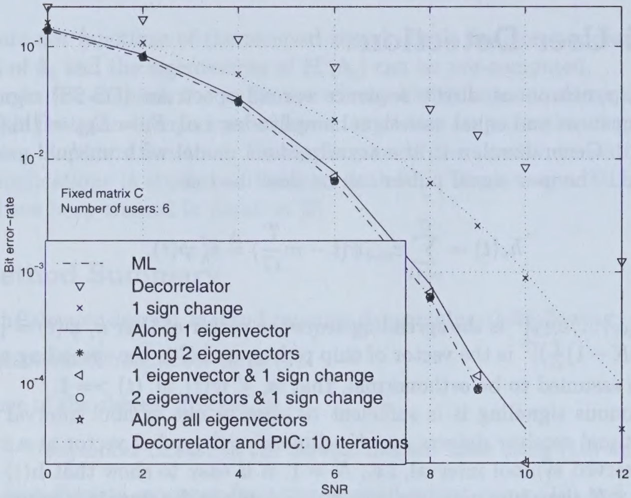


Figure 2: Multi-user detection for a fixed matrix **S**.

between two values as a function of the number of PIC stages. Here we select the better of the two.

Figure 3 shows that when **S** is randomly chosen for each symbol transmission, comparison of admissible symbol vectors along 2 eigenvectors of **Y** allows for practically optimal performance. Nevertheless, the comparison along only 1 eigenvector results in a performance degradation of less than 1.5 dB for bit error-rates of 10^{-4} or larger.

The performances of the decorrelator and that of the method which evaluates the likelihoods of the K closest neighbors (each obtained by changing the sign of a single bit of $\hat{\mathbf{a}}_{dec}$) to $\hat{\mathbf{a}}_{dec}$ are significantly worse. The change-of-sign method [4] has the same complexity as the slowest descent method with $P = 1$ once the sequences closest to $\hat{\mathbf{a}}_c + \rho\hat{\mathbf{\Delta}}_{min}$ have been determined.

5 The ISI Channel

One user, i.e., $K = 1$, known ISI channel, and no additive interference are assumed. Here, an approach to sequence estimation similar to the one described in the previous section for multi-user detection is taken.

The signal pulse autocorrelation matrix is

$$\mathbf{Y} = <\mathbf{h}(t), \mathbf{h}^T(t)> = [<h(t-kT), h(t-mT)>]_{km}. \tag{22}$$

Here

$$\mathbf{H}^l(\mathbf{a}) \equiv -\mathbf{Y} \quad \text{and} \quad \hat{\mathbf{a}}_c = \mathbf{Y}^{-1}\mathbf{r} = \left(<\mathbf{h}(t), \mathbf{h}^T(t)>\right)^{-1} <\mathbf{h}(t), r(t)> \tag{23}$$

is the (non-quantized) output of the zero-forcing equalizer [5]. It is clear that here, as in the case of the decorrelator, by premultiplying the matched-filter output with the inverse channel matrix one eliminates (zero-forces) the intersymbol interference but can enhance noise significantly.

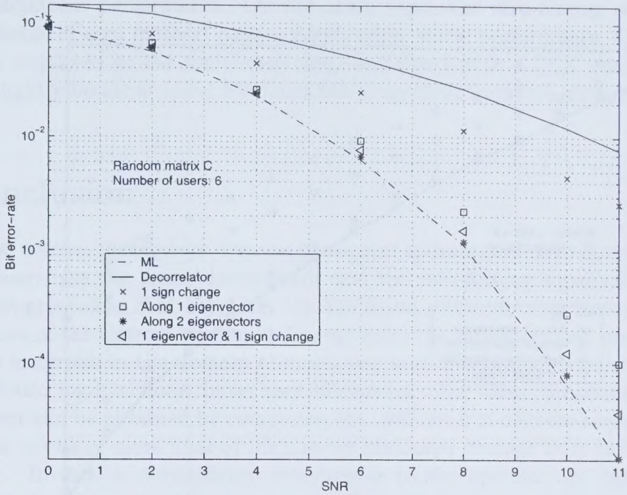


Figure 3: Multi-user detection for a random matrix **S**.

A random discrete channel $\mathbf{h}_d^1 = [h_{d,1}^1, h_{d,2}^1, h_{d,3}^1]^T$ with $L = 3$ taps and a short sequence length $N = 6$ is first analyzed. The channel taps are Gaussian random variables with mean 0 and variance 1/3. The corresponding elements of the autocorrelation matrix **Y** are (see, e.g., [6]):

$$y_{nm} = \sqrt{E_b} \sum_{k=1}^{L-(n-m)} h_{d,k} h_{d,n-m+k} \qquad (24)$$

for $n \geq m$ and $y_{nm} = y_{mn}$ for $n < m$. Using this approach, an average performance over a wide range of channels of length $L = 3$ is assessed. Figure 4 shows that, for a random channel, the comparison of admissible sequences along 2 eigenvectors of **Y** achieves close to optimal performance. The zero-forcing equalizer and the single sign-change method (described in the previous section) have significantly worse performance.

Computation of $\hat{\mathbf{a}}_c$ requires a matrix-vector multiplication with the matrix $\mathbf{Y}^{-1}$ which is not sparse in the general case. This operation has a complexity $O(N^2)$ and can be very expensive for large N. In fact, it can be more expensive in the case of short channels than the Viterbi Algorithm even for relatively short sequences.

Next, we show that the slowest descent method can be applied to a discrete channel of length $L = 10$ where the complexity of the Viterbi Algorithm is significantly larger for either short or moderately long sequences. The employed channel, $\mathbf{h}_d^2 =$ [-0.0465 0.1580 -0.6558 0.2641 -0.3079 -0.2885 0.0891 0.3920 0.3648 -0.0162 $]^T$, has a "deeper" spectral null than the $L = 11$ channel from [6], and, thus, we can expect that the zero-forcing equalizer will not perform as well as for the channel from [6]. The corresponding elements of the autocorrelation matrix **Y** are determined using (24). The number of symbols used in a transmitted block is $N = 100$.

Figure 5 demonstrates that it is sufficient to compare the likelihood of sequences closest to two lines of the slowest descent to obtain a performance close to the non-ISI

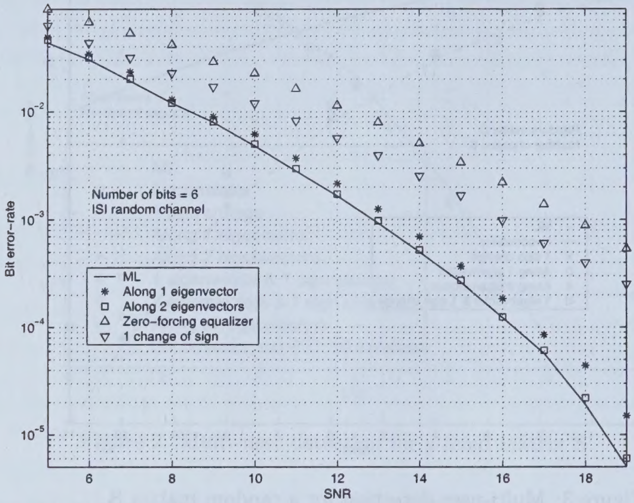


Figure 4: Intersymbol interference channel with random $\mathbf{h}_d$.

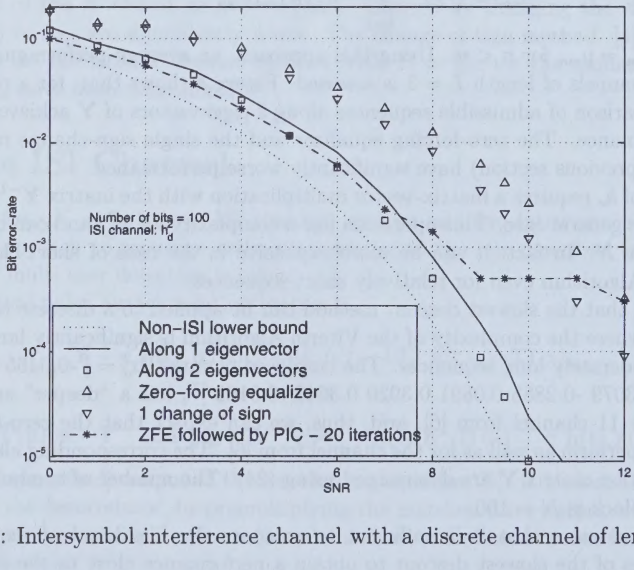


Figure 5: Intersymbol interference channel with a discrete channel of length 10.

lower bound (see, e.g., [6]). Nevertheless, the comparison along only a single descent line gives a comparable performance. On the other hand, the zero-forcing equalizer (ZFE) and the change of sign method have a significantly worse performance. PIC (analogous to the one employed in the multi-user case) initialized with a ZFE using 20 iterations performs slightly better at lower SNR and has a significantly worse performance at larger SNR.

6 Conclusion

The slowest descent method for discrete parameter estimation has been introduced. This method generalizes the linear decorrelator and the zero-forcing equalizer by suggesting further post-processing that compares the likelihood of the discrete sequences along one or more lines of the slowest likelihood descent from the unconstrained maximizer.

Several examples in the domain of multi-user and intersymbol interference detection with quadratic log-likelihood have been addressed. For these problems, a significant improvement can be obtained by comparing the likelihood of discrete sequences closest in the l_2 norm to one or more lines of the slowest likelihood descent from the unconstrained maximizer. In fact, a performance comparable to the optimal one can be obtained. The slowest descent method can also be applied efficiently to problems of detection in the random amplitude channel with either ISI, multi-access interference, or Gaussian interference, that have a non-quadratic log-likelihood function [2].

Acknowledgment

We would like to thank Mr. Qinghua Li for useful discussions that helped in understanding several concepts related to the slowest descent method.

References

[1] G. H. Golub and C. F. V. Loan, *Matrix Computations.* Baltimore, MD: Johns Hopkins University Press, 1989.

[2] P. Spasojević, "Sequence and channel estimation for channels with memory," Ph.D. Dissertation, Texas A&M University, 1999.

[3] R. Lupas and S. Verdú, "Linear multiuser detectors for synchronous code-division multiple-access channels," *IEEE Transactions on Information Theory,* vol. 35, pp. 123–136, January 1989.

[4] Q. Li and C. N. Georghiades, "Approximate ML multi-user detectors for synchronous DS/CDMA channels," in *Proceedings of Conference on Information Sciences and Systems,* Baltimore, MD, pp. 644–649, March 1999.

[5] R. W. Lucky, J. Salz, and E. J. Weldon, *Principles of Data Communication.* New York: McGraw-Hill, 1968.

[6] J. G. Proakis, *Digital Communications.* New York: McGraw-Hill, 1989.

Controllability and Stabilizability of Single Input Planar Juggling

Kevin M. Lynch   Craig K. Black
Mechanical Engineering Department
Northwestern University

Abstract

This paper applies the idea of forced recurrence to demonstrate controllability and stabilizability of a single input juggling system. By the proper choice of a discrete-time control Lyapunov function, nonlinear optimization can be used to find controls in a neighborhood of the recurrent controls that drive the system toward the goal trajectory. The approach is demonstrated on an experimental juggling system.

I. INTRODUCTION AND MOTIVATION

Our work is motivated by the desire to develop motion planning and control algorithms for dynamic systems with enough actuators to be controllable, but not enough to be locally controllable at any state. We are particularly interested in controllable systems with a single input. For example, it is possible to control the attitude dynamics of a satellite (Jurdjevic [1]) or the state of a planar rigid body (Lynch [2]) using a single thruster. The proof of the former example relies on local accessibility and a recurrence property—the unforced tumbling dynamics of the satellite returns the satellite to its initial orientation. In the latter example, the drift motion of the body does not return the system to a neighborhood of its initial state; the same theorems do not apply.

If there exists a state-dependent control law that keeps the system in a recurrent trajectory, however, a similar approach can be taken to proving controllability (Caines and Lemch [3]). In addition, if we can find such a control law, it forms the kernel of a controller to stabilize a desired trajectory.

In this paper we apply these ideas to the problem of controlling a planar juggling system. This system is precisely the system studied by Bühler and Koditschek [4]. It consists of a one-joint revolute arm which bat-juggles a planar puck to a desired vertical limit cycle. Their "mirror law" controller exhibits impressive empirical performance, but the approach has resisted attempts to formally characterize the basin of attraction. While there is a global convergence proof when the configuration of the puck is its height y, there is only a proof of local convergence when the configuration is (x, y).

By utilizing the idea of forced recurrence, we can show controllability on a large computable subset D of the puck's state space, and derive a controller that stabilizes a desired juggling limit cycle from any initial puck state in D. The controller is based on real-time nonlinear optimization with an appropriately chosen discrete-time control Lyapunov function. The robustness of the approach has been demonstrated on an experimental juggling system.

The main purpose of this paper, however, is not to provide a better juggling controller, but to suggest a way to construct controllers for systems which are globally controllable but not locally controllable. This system is our first working example. In future work we hope to apply the approach to other interesting systems, such as the single-input planar body.

In Section II we review the basic ideas behind our approach. In Section III we demonstrate controllability and stabilizability of a simplified juggler, and in Section IV we describe the experimental implementation.

II. THE MAIN IDEA

A. Background: Global Controllability with Drift

Consider a control affine nonlinear system of the form

$$\dot{z} = X_0(z) + \sum_{i=1}^{m} X_i(z)u_i, \quad z \in M, \tag{1}$$

where z is the state, M is a smooth, connected n-dimensional state manifold, X_0 is a drift vector field, $X_1, \ldots X_m$ are control vector fields, and $u = (u_1, \ldots, u_m) \in \mathcal{U}$ is the control input, where $\mathcal{U}$ is a bounded subset of $\Re^m$ containing the origin in the interior of its convex hull. The vector fields $X_0, \ldots, X_m$ are smooth and complete. A trajectory $z(t) : [t_0, t_f] \to M$ is feasible if it is the solution to (1) for some control function $u(t) : [t_0, t_f] \to \mathcal{U}$. The system (1) is (globally) *controllable* on M if z_1 is reachable from z_0 by a feasible trajectory for all $z_0, z_1 \in M$. The system (1) is controllable on a subset $N \subset M$ if z_1 is reachable from z_0 for all $z_0, z_1 \in N$. The trajectories need not remain in N.

In the next three subsections we briefly review results needed to conclude global controllability of certain systems of the form (1).

A.1 Accessibility

Define $R^{\mathcal{B}(z)}(z, T)$ to be the reachable set from z at time $T > 0$ by feasible trajectories remaining in the neighborhood $\mathcal{B}(z)$ of z, and $R^{\mathcal{B}(z)}(z, \leq T) = \bigcup_{0 \leq t \leq T} R^{\mathcal{B}(z)}(z, t)$. Then the system (1) is small-time locally accessible from z if $R^{\mathcal{B}(z)}(z, \leq T)$ contains a nonempty open set of M for all neighborhoods $\mathcal{B}(z)$ and all $T > 0$. The system is small-time locally accessible if this holds for all $z \in M$.

Small-time local accessibility of the system (1) can be established by an algebraic test on the system vector fields. Let $\mathcal{X}(M)$ denote the algebra of all smooth vector fields on M. The accessibility Lie algebra $\mathcal{L}$ generated by the vector fields $X_0, X_1, \ldots, X_m$ is the smallest subalgebra of $\mathcal{X}(M)$ containing $X_0, X_1, \ldots, X_m$. The tangent vectors of vector fields in $\mathcal{L}$ evaluated at z are written $\mathcal{L}(z)$.

Definition 1: The system (1) satisfies the Lie algebra rank condition (LARC) at z if $\mathcal{L}(z) = T_z M$. The system is said to satisfy the LARC if it is satisfied at all $z \in M$.

If the LARC is satisfied, the system is small-time locally accessible. In the special case that $X_0 = 0$ for the system (1), the LARC implies small-time local controllability, that is, $R^{\mathcal{B}(z)}(z, \leq T)$ contains a neighborhood of z for all $z \in M, T > 0$ and all neighborhoods $\mathcal{B}(z)$ of z. A more general sufficient condition for small-time local controllability is given by Sussmann [5]. The systems we consider in this paper are not small-time locally controllable.

A.2 Weakly Positive Poisson Stability

Let $\phi^{X_0} : M \times \Re \to M$ denote the flow of the drift vector field X_0, so that $\phi^{X_0}(z, t)$ gives the system state after following the flow ϕ^{X_0} from z for a time t. A point z in M is called positively Poisson stable for X_0 if for all $T > 0$ and any neighborhood U_z of z, there exists a time $t > T$ such that $\phi^{X_0}(z, t) \in U_z$. The vector field X_0 is called positively Poisson stable if the set of positively Poisson stable points for X_0 is dense in M.

A point $z \in M$ is called a nonwandering point of X_0 if for all time $T > 0$ and any neighborhood U_z of z there exists a time $t > T$ such that $\phi^{X_0}(U_z, t) \cap U_z \neq \emptyset$, where $\phi^{X_0}(U_z, t) = \{\phi^{X_0}(z, t) | z \in U_z\}$. (A positively Poisson stable point is necessarily a nonwandering point.) The nonwandering set of X_0 is the set of all nonwandering points of X_0.

Definition 2: The vector field X_0 is weakly positively Poisson stable (WPPS) if its nonwandering set is M.

Recurrent drift vector fields are an important subset of WPPS drift vector fields. In this paper we use the following definition.

Definition 3: A point z is recurrent for X_0 if $\phi^{X_0}(z, t) = z$ at some time $t > T$ for any $T > 0$. X_0 is called recurrent if the set of recurrent points is dense.

A.3 Global Controllability

We now state the main theorem, taken from Lian *et al.* [6]. Related results can be found in (Jurdjevic and Sussmann [7]; Lobry [8]; Brockett [9]; Jurdjevic [1]).

Theorem 1: Assume that the drift vector field X_0 is weakly positively Poisson stable (WPPS). Then the system (1) is controllable on M if the Lie algebra rank condition (LARC) is satisfied.

Theorem 1 can be used to demonstrate the velocity controllability of a planar body with a single jet (Manikonda and Krishnaprasad [10]), controllability of the attitude dynamics of a satellite with a single thruster (Jurdjevic [1]), or the controllability of satellite orbital dynamics through the control of attitude dynamics (Lian *et al.* [6]).

B. Systems without WPPS Drift

Theorem 1 is a powerful tool for establishing global controllability of underactuated systems with drift, particularly those evolving on compact manifolds. For example, if the drift field is phase-volume conserving and the state space is compact, the drift field is WPPS (Arnold [11]; Lian *et al.* [6]). (See also [10] for a condition for noncompact state spaces.) In many situations, however, the drift field is not WPPS. In the juggling system, for instance, the puck does not repeatedly return to a neighborhood of its initial state in the absence of control.

Similar arguments can be applied, however, if it is possible to impose a state feedback control on the system such that all points $z \in M$ are nonwandering. In particular, we can define the notion of *control recurrence* (Caines and Lemch [3]). Let $\phi(z, t, u)$ denote the flow of (1) under the control u.

Definition 4: A state $z \in M$ is *control recurrent* if there exists a control $u : [0, t] \to \mathcal{U}$ such that $z = \phi(z, t, u)$ at some time $t > T$ for any $T > 0$.

In this paper we will find it useful to define a stronger version of control recurrence.

Definition 5: A state $z \in M$ is *strongly control recurrent* if there exists a control $u : [0, t] \to \text{int}(\mathcal{U})$ such that $z = \phi(z, t, u)$ at some time $t > T$ for any $T > 0$.

The set R is called control recurrent if each state $z \in R$ is control recurrent. Let $W \subseteq R$ be the set of control recurrent states such that there exists a recurrent trajectory about which the system is controllable. Then the system is controllable on any connected subset of W.

C. Controllability and Control of Discrete-Time Systems

The juggling system is a discrete-time control system, as control of the puck's motion is only exerted at instants of impact. In this section we adapt the ideas above to discrete-time systems. We refer the reader to [12] for a review of results on controllability of discrete-time nonlinear control systems.

As before, z is a point in the smooth, connected n-dimensional state manifold M. Define a k-dimensional control vector $u = (u_1, \ldots, u_k)^T \in \mathcal{U}$, where $\mathcal{U}$ is the set of feasible controls. The discrete-time nonlinear control system can be written

$$z_{j+l} = f_l(z_j, u) \tag{2}$$

where $f_l : M \times \mathcal{U} \to M$ represents the system dynamics mapping an initial state and a control vector to a final state. Note that f_l represents the dynamics over l control events; the states z_j and z_{j+l} are consecutive if $l = 1$. Clearly $\dim(\mathcal{U})$ increases as l increases. In our formulation of f_l we assume l has been chosen large enough such that $\dim(\mathcal{U}) \geq \dim(M)$. We require that f_l be at least once-differentiable with respect to u; that is, the Jacobian $\frac{\partial f_l(z,u)}{\partial u}$ exists.

For the discrete-time system (2), we can define control recurrence as follows.

Definition 6: A state z is discrete-time control recurrent if there exists $u \in \mathcal{U}$ such that $z = f_l(z, u)$. A state z is discrete-time strongly control recurrent if there exists $u \in \text{int}(\mathcal{U})$ such that $z = f_l(z, u)$. Any control u satisfying the condition for (strong) control recurrence at z is called a (strongly) recurrent control for z.

A discrete-time control recurrent state may also be viewed as a forced equilibrium. For the rest of the paper, we will omit the qualifier "discrete-time" and regard it as understood.

Let $R \subseteq M$ be the set of strongly control recurrent states. Let $D \subseteq R$ be a p-dimensional open connected subset of R such that there is a known (and computable) strongly recurrent control mapping $u^r : D \to \text{int}(\mathcal{U})$. Repeated execution of the control $u^r(z)$ for any $z \in D$ keeps the system in a single recurrent orbit.

With these definitions, a sufficient condition for controllability on D is that a neighborhood of the control $u^r(z)$ for all $z \in D$ maps through the system dynamics to a neighborhood of $z = f_l(z, u^r(z))$. This assures local controllability about all states $z \in D$. A sufficient condition for local controllability is linear controllability about the recurrent orbit:

$$\text{rank } \frac{\partial f_l(z,u)}{\partial u}|_{u=u^r(z)} = p \quad \forall z \in D.$$

Our ultimate goal is to derive a controller to stabilize the system to a goal state $z_g \in D$. In this paper, we assume a metric on states in D, $d : D \times D \rightarrow \Re$. Each state $z \in D$ has an associated cost $V(z) = d(z, z_g)$, where $z_g \in D$ is the goal control recurrent state. We define $\nabla_u V(z) = \frac{\partial V(f(z,u))}{\partial u} = \frac{\partial V}{\partial z}\frac{\partial f}{\partial u}$, the gradient of the cost with respect to the control parameters. This gradient is typically evaluated at the recurrent control, $\nabla_u V(z)|_{u=u^r(z)}$. If V is positive definite and $\nabla_u V(z)|_{u=u^r(z)} \neq 0$ for all $z \in D$ except at the desired state z_g, then V acts as a discrete-time control Lyapunov function, and the following algorithm suffices to achieve asymptotic convergence to z_g from any state in D:

1. Calculate the recurrent control $u^r(z_0)$ for the current state z_0.
2. Perform a single step of a gradient descent algorithm using $\nabla_u V(z)|_{z=z_0, u=u^r(z_0)}$ to find a feasible control u' such that $V(f(z_0, u')) < V(f(z_0, u^r(z_0))) = V(z_0)$ and $f(z_0, u') \in D$.
3. Execute the control u' and return to step 1.

It is assumed that calculation of the recurrent control and the gradient takes constant time. Usually the gradient step requires a line-search algorithm, the execution time of which can be bounded based on the required precision of the search. Thus control computation time can be bounded, allowing the possibility of real-time control.[1]

In practice, the approach outlined above may be difficult to implement. It requires a known differentiable forward dynamics model f_l, a positive definite V with no local extrema, and the ability to calculate a recurrent control.

For the problem of juggling a puck, however, the formulation above allows the demonstration of controllability on a large and computable subset D of the state space and the derivation of a controller which asymptotically stabilizes the system to the desired juggling limit cycle from any point in D. In this case $l = 2$, that is, the dynamics f_2 are defined over two bats. The control law is called the two-bat control law (2BCL).

For practical reasons, we seek a one-bat control law (1BCL) for the system f_1, consisting of a single bat. In this case, each bat is the control that reverses the trajectory, plus a step in the direction $\nabla_u V$. For f_1, however, there are fewer controls than the dimension of the state space, and therefore we may have that $\nabla_u V = 0$ at some states. However, this condition cannot persist on consecutive states (except at the goal state) for an appropriate choice of V. By an application of the invariance principle, we find that the controller aysmptotically stabilizes the goal state with the domain of attraction D. We have implemented a variation of the 1BCL on our juggling apparatus and have empirically verified its robust stabilization properties.

III. Analysis of a Simplified Juggler

A. Juggler Dynamics

The one-joint revolute robot arm is a half-line with a pivot at the endpoint. A world frame $\mathcal{F}_W$ is fixed at the pivot, with the gravitational acceleration $g < 0$ acting in the $-y$ direction (see Figure 1). The arm angle θ with respect to the x-axis of $\mathcal{F}_W$ is restricted to the range $(-\pi/2, \pi/2)$. The configuration of the point mass puck in $\mathcal{F}_W$ is $q = (x, y)^T \in \mathcal{C} = \Re^2$. The state of the puck is $z = (q^T, \dot{q}^T)^T \in T\mathcal{C}$,

---

[1] The basic approach can also be applied to continuous-time systems by considering the dynamics on a Poincaré section, choosing a finite parameterization of the control $u(t)$, and defining a cost function on states on the Poincaré section such that the gradient of the cost with respect to the control parameters about a computable recurrent control is nonzero everywhere except at the goal state.

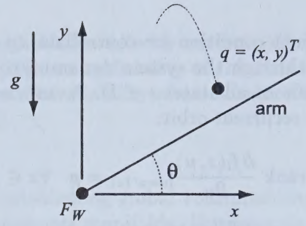


Fig. 1. Notation for the simplified juggler.

and we have the natural projection $\pi : T\mathcal{C} \to \mathcal{C}$. We define $rev(z) = (q^T, -\dot{q}^T)^T$, where the velocity reversed. For analysis we treat the position θ and velocity $\dot{\theta}$ of the arm as control inputs, so the sta of the system is simply z.

The puck's motion during free flight is described by the mapping $a : T\mathcal{C} \times \Re \to T\mathcal{C}$

$$a(z_0, t) = (x_0 + \dot{x}_0 t,\ y_0 + \dot{y}_0 t + g\frac{t^2}{2},\ \dot{x}_0,\ \dot{y}_0 + gt)^T .$$

The path of the puck during flight is a parabola. For any state $z = (x, y, \dot{x}, \dot{y})^T$ there is a uniq associated free flight trajectory with a unique apex point where the puck attains its maximum y valu $apex(z) = (x - \dot{x}\dot{y}/g,\ y - \dot{y}^2/2g,\ \dot{x},\ 0)^T$. If $\dot{y} > 0$, then $apex(z)$ lies on the forward time flow of a, an if $\dot{y} < 0$, $apex(z)$ lies on the reverse time flow. Let $A = \{z|\dot{y} = 0\}$ denote the three-dimensional s of apex states. A parameterizes the space of free-flight parabolas, and $(x, y, \dot{x}, 0)^T$ and $(x, y, -\dot{x}, 0)$ correspond to the same parabola.

An impact mapping $b : T\mathcal{C} \times \Re \to T\mathcal{C}$ takes a state z^- immediately before impact, along with th arm velocity $\dot{\theta}$ at impact, and produces the state z^+ immediately after impact. For ease of derivatio we write

$$b(z^-, \dot{\theta}) = z^+ = (x, y, (b_3(b_2(b_1(\dot{q}^-, \dot{\theta})), \dot{\theta}))^T)^T,$$

where b_1 maps the pre-impact velocity $\dot{q}^-$ and $\dot{\theta}$ to puck impact velocity v^- in an arm-fixed impa frame $\mathcal{F}_I$, b_2 maps pre-impact to post-impact velocity in $\mathcal{F}_I$, and b_3 maps the post-impact velocity from $\mathcal{F}_I$ back to $\mathcal{F}_W$.

To derive b_1 and b_3, let $\mathcal{F}_I$ be a frame attached to the arm at the impact point $q = (x, y)^T$, wit unit normal and radial directions (relative to the arm pivot)

$$\hat{n} = \frac{(-y, x)^T}{\sqrt{x^2 + y^2}}, \quad \hat{r} = \frac{(x, y)^T}{\sqrt{x^2 + y^2}}.$$

The pre-impact velocity of the puck in $\mathcal{F}_I$ is

$$b_1(\dot{q}^-, \dot{\theta}) = v^- = (v_r^-, v_n^-)^T = ((\dot{q}^-)^T\hat{r},\ (\dot{q}^-)^T\hat{n} - \dot{\theta}\sqrt{x^2 + y^2})^T .$$

The conditions for an impact are $\theta = \tan^{-1}(y/x), v_n^- < 0$. A post-impact velocity v^+ in $\mathcal{F}_I$ is transformed to $\mathcal{F}_W$ by

$$b_3(v^+, \dot{\theta}) = \dot{q}^+ = (v_r^+)\hat{r} + (v_n^+ + \dot{\theta}\sqrt{x^2 + y^2})\hat{n}.$$

b_2 is derived from Newton's restitution law:

$$b_2(v^-) = v^+ = (v_r^-, -ev_n^-)^T,$$

where $e \in [0,1]$ is the *coefficient of restitution*. The arm is assumed infinitely massive relative to the puck, so that the velocity of the arm does not change at impact.

Let $u_1 = (t_1, \dot{\theta}_1, t_f)^T \in U_1 = \Re^+ \times \Re \times \Re^+$ define the control set for batting the puck once, where $\Re^+$ denotes the set of positive reals. The puck flies for time t_1, is batted with arm velocity $\dot{\theta}_1$, and flies for time t_f. We define the discrete-time forward dynamics $f_1 : TC \times U_1 \to TC$ for a single bat

$$f_1(z_0, u_1) = a(b(a(z_0, t_1), \dot{\theta}_1), t_f). \tag{8}$$

Similarly we can define the forward dynamics $f_2 : TC \times U_2 \to TC$ for two bats

$$f_2(z_0, u_2) = a(b(a(b(a(z_0, t_1), \dot{\theta}_1), t_2), \dot{\theta}_2), t_f), \tag{9}$$

summarized as flight for t_1, bat with $\dot{\theta}_1$, flight for t_2, bat with $\dot{\theta}_2$, and flight for t_f, and $u_2 = (t_1, \dot{\theta}_1, t_2, \dot{\theta}_2, t_f)^T \in U_2 = \Re^+ \times \Re \times \Re^+ \times \Re \times \Re^+$. Note that U_1 and U_2 are both open sets. Also, t_f is not a control in the usual sense, in that it does not affect the trajectory of the puck; it simply chooses the state used to represent the final free-flight trajectory.

B. Reversible Trajectories

For the juggling system, we say that a state $z = (q^T, \dot{q}^T)^T$ is *reversible* if there exists a control $u_2^r = (t_1^r, \dot{\theta}_1^r, t_2^r, \dot{\theta}_2^r, t_f^r)^T$ such that the puck attains the state $rev(z)$ during free-flight after the first bat and the original state z after the second bat ($z = f_2(z, u_2^r)$). Reversible states are a subset of strongly control recurrent states.

Lemma 1: Let $z = (x, y, \dot{x}, \dot{y})^T$ and τ_1, τ_2, τ_3 be the solutions for t of the cubic equation

$$(x + \dot{x}t)\dot{x} + (y + \dot{y}t + gt^2/2)(\dot{y} + gt) = 0 \tag{10}$$

called the *normal impact condition*. Then z is reversible if and only if (1) there exists a solution τ_1 such that $\tau_1 \in \Re, \tau_1 > 0, \dot{y} + g\tau_1 < 0, x + \dot{x}\tau_1 > 0$, and (2) there exists a solution τ_2 such that $\tau_2 \in \Re, \tau_2 < 0, \dot{y} + g\tau_2 > 0, x + \dot{x}\tau_2 > 0$.

Proof: Omitted for space.

Remark: We can interpret Equation (10) geometrically to gain intuition. Consider the configuration space parabola described by $a(z,t)$ as t varies from $-\infty$ to ∞. The points on the parabola at which the normal impact condition is satisfied are those where (1) the puck is at the origin or its velocity is zero (only at the apex of vertical trajectories) or (2) the parabola is tangent to a circle centered at the origin of $\mathcal{F}_W$. Case (1) is not of interest; we focus only on case (2) solutions. At these points, the velocity of the puck is normal to the line connecting the puck to the origin (the arm). Figure 2 shows example parabolic flights, reversible states, and the points on the parabolas corresponding to case (2) solutions. Because (10) is cubic in t, there are at most three real solutions.

Lemma 2 (Reversing control.) Let τ_1, τ_2, τ_3 be the three solutions of t for Equation (10). If z is reversible by Lemma 1, then all three solutions are real, and $\tau_{max} = \max\{\tau_1, \tau_2, \tau_3\} > 0, \tau_{min} = \min\{\tau_1, \tau_2, \tau_3\} < 0$. Then $u_2^r(z) = (t_1^r, \dot{\theta}_1^r, t_2^r, \dot{\theta}_2^r, t_f^r)^T$ is a reversing control if

$$t_1^r = \tau_{max},\ t_2^r = \tau_{max} - \tau_{min},\ t_f^r = -\tau_{min},\ \dot{\theta}_1^r = \frac{(1-e)\sqrt{\dot{x}_1^2 + \dot{y}_1^2}}{(1+e)\sqrt{x_1^2 + y_1^2}},\ \dot{\theta}_2^r = \frac{(1-e)\sqrt{\dot{x}_2^2 + \dot{y}_2^2}}{(1+e)\sqrt{x_2^2 + y_2^2}},$$

where $z_1 = a(z, \tau_{max})$ and $z_2 = a(rev(z), -\tau_{min})$, the states immediately before the first and second impacts.

To visualize the set of reversible states, we can look at the apex states of their corresponding parabolas.

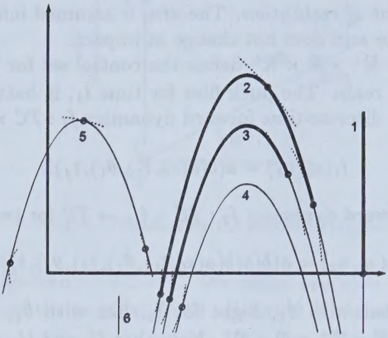


Fig. 2. **Examples of reversible states. Shown are six example parabolic flights, where each point on a parabola represents two states of the puck, corresponding to the two possible directions of travel along the parabola. (At the apex of a vertical trajectory, these are the same state, with $\dot{x} = \dot{y} = 0$.) Shown in bold are reversible states, and the dots indicate the impact points where the puck velocity can be reversed for $\dot{y} < 0$. Also shown are arcs of the circles centered at the origin and tangent to the parabolas at these points. (Parabola 1) States in a vertical trajectory are reversible by impact at $y = 0$. (Parabola 2) The typical case, there are three distinct real solutions to the normal impact condition (10). (Parabola 3) This is a copy of Parabola 2 with the apex lowered. Here, the two distinct real solutions on the right side of the parabola have collapsed into the same solution. (Parabola 4) This is the copy of Parabola 3 with the apex lowered. The real solutions on the right side have become complex. No state is reversible; there is only one real solution to (10). (Parabola 5) Although there are three real solutions, no state is reversible, as the arm cannot bat the puck at the leftmost solution point. (Parabola 6) The vertical trajectory contains no points with $y \geq 0$, so the velocity cannot be reversed at any state on the trajectory.**

Lemma 3: An apex state $z = (x, y, \dot{x}, 0)^T$ is reversible if and only if $x > 0, y > 0$ and

$$\frac{-gx^2}{2|\dot{x}|} > y \geq \frac{-3(x|\dot{x}|g)^{2/3} - 2\dot{x}^2}{2g}. \tag{11}$$

Remark: By the definition of a reversible state, and by the conditions of the lemma, it is clear that if $(x, y, \dot{x}, 0)^T$ is reversible, then so is $(x, y, -\dot{x}, 0)^T$.

Proof: Omitted for space.

Figure 3 illustrates the set of reversible apex states. This set contains boundary states defined by the lower bound of y in (11). We define D_A to be the open set of reversible apex states obtained by eliminating these boundary states. D_A is an open connected subset of A, and it contains all states of the form $(x, y, 0, 0)^T$ where $x, y > 0$. Then $D = \{a(z, \tau) | z \in D_A, \tau \in (\tau_{min}(z), \tau_{max}(z))\}$ is a connected four-dimensional open set of reversible states on $\mathcal{TC}$ obtained by the mapping of the open time interval (between the extremal real solutions of (10)) through a for each $z \in D_A$.

C. Controllability

Our next task is to establish controllability of the juggling system on D. This comes from linear controllability of $f_2(z, u_2), z \in D$ at the reversing control $u_2^r(z)$.

Lemma 4: For all $z \in D$, rank $\frac{\partial f_2(z,u_2)}{\partial u_2}|_{u_2 = u_2^r(z)} = 4$.

The proof is based on a detailed computation using Mathematica, which is omitted.

Lemma 4 leads easily to the following proposition.

Proposition 1: The point mass puck is controllable from any $z_1 \in D$ to any other $z_2 \in D$.

Proof: At any state $z \in D$, we have $z \in \text{int}(f_2(z, \mathcal{B}(u_2^r(z))))$, where $f_2(z, \mathcal{B}(u_2^r(z))) = \{f_2(z, u_2) | u_2 \in \mathcal{B}(u_2^r(z))\}$ and $\mathcal{B}(u_2^r(z))$ is any open set of U_2 containing $u_2^r(z)$. This follows from Lemma 4. Further, since U_2 is open, there is a feasible neighborhood of controls of $u_2^r(z)$.

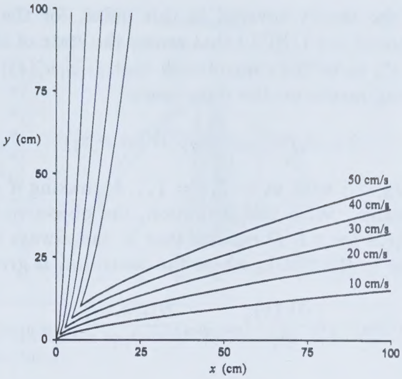


Fig. 3. The set of reversible apex states $z = (x, y, \dot{x}, 0)^T$ plotted at constant $|\dot{x}|$ slices with $g = -981\text{cm/s}^2$. For an apex velocity $\dot{x} = 0$, the region includes the entire open top right quadrant $x, y > 0$. The region shrinks as $|\dot{x}|$ increases. For a different value of gravity g', the velocities at the slices shown are scaled by $\sqrt{g'/g}$.

Now consider any curve γ in D connecting z_1 to z_2. Such a curve exists because D is connected, and every point on γ is in the interior of D. By the property above, every point z on γ is in the interior of the set accessible from z by two bats. These sets form an open cover of the curve, and because the curve is compact, there is a finite subcover. z_2 can be reached from z_1 in a finite number of bats. □

D. Stabilizing Controllers

In this paper the goal of the controller is to stabilize a vertical juggling trajectory with an apex state $z_g = (x_g, y_g, 0, 0)^T \in D$. With the metric $d(z_1, z_2)$, we define the objective function (also the candidate discrete-time control Lyapunov function) $V(z) = d(z, z_g)$. The purpose of the controller is to drive $V(z)$ to zero. We choose d such that $V(z)$ is smooth, positive definite over D, and $\frac{\partial V}{\partial z} \neq 0$ except at $z = z_g$.

Following the basic algorithm in Section II-C, the two-bat control law (2BCL) for (9) is: Calculate the reversing control $u_2^r(z_0) = (t_1^r, \dot{\theta}_1^r, t_2^r, \dot{\theta}_2^r, t_f^r)^T$ for the current state z_0 (which keeps the objective function constant), follow the gradient $\nabla_{u_2} g(f_2(z, u_2))$ evaluated at $(z_0, u_2^r(z_0))$ to find a control u_2' which decreases the objective function and satisfies $f_2(z_0, u_2') \in D$ (states after the first bat must also be in D), and execute the control u_2'. In its basic form, the 2BCL executes two bats between sensing events.

Lemma 5: At any $z_0 \in D, z_0 \neq z_g$, $\nabla_{u_2} V(f_2(z, u_2))|_{z=z_0, u_2=u_2^r(z_0)} \neq 0$.

Proof: We have $\nabla_{u_2} V(f_2(z, u_2)) = \frac{\partial V}{\partial z} \frac{\partial f_2}{\partial u_2}$. At $z_0 \neq z_g$, by assumption we have $\frac{\partial V}{\partial z} \neq 0$, and by Lemma 4 we have full rank of $\frac{\partial f_2}{\partial u_2}|_{u_2=u_2^r(z)}$. Therefore, the gradient of V is nonzero. □

Lemma 5 leads easily to the desired result.

Proposition 2: Under the control law 2BCL, the system (9) asymptotically converges from any $z \in D$ to the desired vertical juggling limit cycle with the apex state $z_g = (x_g, y_g, 0, 0)^T \in D$.

Proof: For any $z_0 \in D$, consider the control

$$u_2' = u_2^r(z_0) - \alpha \nabla_{u_2} V(f_2(z, u_2))|_{z=z_0, u_2=u_2^r(z_0)}.$$

Because the gradient $\nabla_{u_2} V$ is nonzero for all $z \in D, z \neq z_g, u_2 = u_2^r(z)$, and because D is open, there exists a sufficiently small $\alpha > 0$ such that $V(f_2(z_0, u_2')) \in D$ and $V(f_2(z_0, u_2')) < V(f_2(z_0, u_2^r(z)))$ at all $z \neq z_g$. □

While the 2BCL embodies the theory covered to this point, for the juggling system it is more practical to define a one-bat control law (1BCL) that senses the state of the system between each bat. We define $u_1^r(z) = (t_1^r, \dot{\theta}_1^r, t_f^r = t_1^r)$ to be the control such that $f_1(z, u_1^r(z)) = rev(z)$ for all $z \in D$. For the 1BCL we define the following metric on the state space:

$$d(z_1, z_2) = (z_1 - z_2)^T W (z_1 - z_2).$$

We choose $W = \mathrm{diag}(w_1, w_2, w_3, w_4)$ with $w_i > 0, i = 1 \ldots 4$, making d a positive definite quadratic function of the difference of states. With this definition, the objective function is $V(z) = d(z, z_g)$. The existence of the control $u_1^r(z)$ for $z \in D$ ensures that V can always be kept constant, as $V(z) = V(rev(z))$. The 1BCL is similar to the 2BCL, where the control u_1' is given by

$$u_1' = u_1^r(z_0) - \alpha \frac{\partial V(z)}{\partial z}\Big|_{z=rev(z_0)} \frac{\partial f_1(z, u_1)}{\partial u_1}\Big|_{z=z_0, u_1=u_1^r(z_0)}$$

for a sufficiently small $\alpha > 0$.

Because u_1 consists of only three controls, however, clearly $\frac{\partial f_1}{\partial u_1}$ is not full rank. When the null space of this matrix contains $\nabla_{u_1} V(z)|_{z=rev(z)}$, the normal to the hyperellipsoids defined by W, then $\nabla_{u_1} V = 0$ at the reversing control. To first-order, small variations δu_1 about $u_1^r(z)$ do not affect V. However, the key insight is that this situation cannot occur on consecutive bats on a reversing trajectory; therefore, it is always possible to decrease the objective function after two bats. This is an application of a discrete version of the invariance principle; the only set of states which is invariant under the 1BCL is $z = z_g$.

Lemma 6: For all $z_0 \in D, z \neq z_g$, if $\nabla_{u_1} V(f_1(z, u_1))|_{z=z_0, u_1=u_1^r(z_0)} = 0$, then $\nabla_{u_1} V(f_1(z, u_1))|_{z=rev(z_0), u_1=u_1^r(rev(z_0))} \neq 0$.

Proof: Omitted for space.

Proposition 3: Under the control law 1BCL, the system (8) asymptotically converges from any $z \in D$ to the desired vertical juggling limit cycle with the apex state $z_g = (x_g, y_g, 0, 0)^T \in D$.

In the next section we describe our implemented variation of the 1BCL and experimentally verify its robust convergence properties. In practice we use the Hessian $\nabla^2_{u_1 u_1} V$ to improve our search for the control u_1', and we take more than a single Newton step during each control computation. Because V, f_1, and their derivatives are smooth, the performance of the controller degrades gracefully with state measurement error, a crucial requirement for reliable experimental performance.

IV. IMPLEMENTATION

FLATLAND is an air hockey table which floats planar parts with low friction. The table is mounted on a pivoting frame, which allows us to simulate zero gravity or nearly full gravity. In our experiments, the table is tilted 5 degrees from horizontal, which increases the time-scale of manipulation by a factor of approximately 3.4 compared to full gravity. A one-joint revolute robot powered by an HD Systems harmonic drive actuator with a 100,000 pulse/rev encoder is mounted to the side of the table. The robot is controlled by a 300 MHz Pentium II PC at a servo rate of 1 kHz. An overhead camera tracks colored blobs affixed to the puck. Vision data is provided at 60 Hz by the Cognachrome vision system. We use the last twenty data points (1/3 sec) to estimate the current position and velocity of the puck.

In our experiments we use a plastic circular puck with a radius of 3.8cm. The batting arm is 60cm long and 5cm wide. We adopt Wang and Mason's[13] planar impact equations for the implementation of the 1BCL. This model uses the Poisson restitution coefficient e and the Coulomb friction coefficient μ, which are approximately 0.45 and 0.1 for our system, respectively. These are the primary differences in the simplified model studied in the previous section and our actual system. As the width of the arm and the puck go to zero, the results produced by Wang and Mason's model approach those given by the Newton impact model.

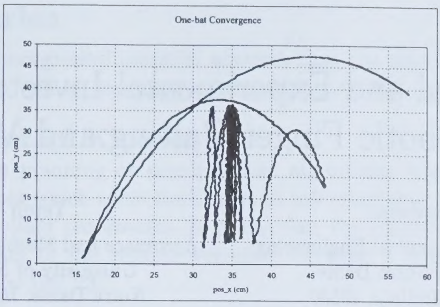


Fig. 4. Raw vision data showing convergence to a stable batting cycle under the control law 1BCL. The desired apex of the puck is $(x_g, y_g) = (35cm, 35cm)$.

We use a BFGS quasi-Newton optimization, and the weight matrix is $W = \mathrm{diag}(4, 4, 0.5, 0.5)$. The controls are implemented by following a a fourth-order polynomial trajectory taking the arm to the desired impact state. We continually re-optimize the solution as new state data become available every 1/60 second. More details on the implementation can be found in (Lynch and Black [14]).

Figure 4 shows an experimental result with $(x_g, y_g) = (35\text{cm}, 35\text{cm})$. This trial is representative of the large domain of attraction that we observed empirically.

To test the robustness of the controller, we dropped the puck ten times from each of four different initial configurations, $(x_i, y_i) = (\{15cm, 27cm, 39cm, 51cm\}, 53cm)$. The system achieved the stable batting cycle 37 of the 40 times. Because the optimal control is continually updated using new sensor data, juggling is robust to perturbations such as bumping the puck or temporarily stopping its motion. Video of the juggler is at `http://lims.mech.northwestern.edu/~lynch/research/videos`.

ACKNOWLEDGMENTS

This work was supported by NSF grant IIS-9811571.

REFERENCES

[1] Velimir Jurdjevic, *Geometric Control Theory*, Cambridge University Press, 1997.

[2] Kevin M. Lynch, "Controllability of a planar body with unilateral thrusters," *IEEE Transactions on Automatic Control*, vol. 44, no. 6, pp. 1206–1211, June 1999.

[3] Peter E. Caines and Ekaterina S. Lemch, "On the global controllability of nonlinear systems: Fountains, recurrence, and applications to Hamiltonian systems," 1999, Submitted to *SIAM J. Cont. Opt.*

[4] M. Bühler and D. E. Koditschek, "From stable to chaotic juggling: Theory, simulation, and experiments," in *IEEE International Conference on Robotics and Automation*, Cincinnati, OH, 1990, pp. 1976–1981.

[5] Héctor J. Sussmann, "A general theorem on local controllability," *SIAM Journal on Control and Optimization*, vol. 25, no. 1, pp. 158–194, Jan. 1987.

[6] K. Lian, L. Wang, and L. Fu, "Controllability of spacecraft systems in a central gravitational field," *IEEE Transactions on Automatic Control*, vol. 39, no. 12, pp. 2426–2440, Dec. 1994.

[7] Velimir Jurdjevic and Héctor J. Sussmann, "Control systems on Lie groups," *Journal of Differential Equations*, vol. 12, pp. 313–329, 1972.

[8] Claude Lobry, "Controllability of nonlinear systems on compact manifolds," *SIAM Journal on Control*, vol. 12, no. 1, pp. 1–4, 1974.

[9] R. W. Brockett, "Nonlinear systems and differential geometry," *Proceedings of the IEEE*, vol. 64, no. 1, pp. 61–72, Jan. 1976.

[10] Vikram Manikonda and P. S. Krishnaprasad, "Controllability of Lie-Poisson reduced dynamics," in *American Control Conference*, 1997.

[11] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, Springer-Verlag, 1989.

[12] Bronislaw Jakubczyk and Eduardo D. Sontag, "Controllability of nonlinear discrete-time systems: A Lie algebraic approach," *SIAM Journal on Control and Optimization*, vol. 28, no. 1, pp. 1–33, Jan. 1990.

[13] Yu Wang and Matthew T. Mason, "Two-dimensional rigid-body collisions with friction," *ASME Journal of Applied Mechanics*, vol. 59, pp. 635–641, Sept. 1992.

[14] Kevin M. Lynch and Craig K. Black, "Control of underactuated manipulation by real-time nonlinear optimization," in *International Symposium on Robotics Research*. 1999, MIT Press.

Theoretical and Experimental Investigation of Stratified Robotic Finger Gaiting and Manipulation*

Bill Goodwine
Aerospace and Mechanical Engineering
University of Notre Dame
Notre Dame, Indiana 46556
goodwine@controls.ame.nd.edu

Yejun Wei
Aerospace and Mechanical Engineering
University of Notre Dame
Notre Dame, Indiana 46556
ywei@nd.edu

Abstract

This paper presents the application of stratified motion planning to the robotic manipulation problem. Although the manipulation problem is a subclass of applications for stratified motion planning, the method present is general in that it is formulation in a manner independent of the object surface geometry or the kinematics of the "fingers" of the robot. The theoretical development of the method is presented as well as experimental results.

1 Introduction

This paper presents the development and experimental verification of a general control framework for robotic grasping and manipulation problems where "fingers" manipulate a grasped object. The method incorporates standard techniques from nonlinear control; furthermore, it *analytically* incorporates techniques to exploit the discontinuities present if the fingers intermittently contact the object (such manipulation has been called "finger gaiting"). Incorporating the discontinuities of the equations of motion of a system into a general motion planning algorithm is difficult because almost all motion planning methods assume that the equations of motion are smooth.

Robotic grasping and manipulation have been the subject of many research efforts, and only an overview can be provided here. Vast efforts have been directed toward the *analysis* of grasp stability and force closure [28, 29, 32], motion planning assuming continuous contact [22, 37, 13] and haptic interfaces and other sensing [4, 31, 30]. *Finger gaiting*, where fingers make and break contact with the object has been less extensively considered and is the main focus of this paper. Finger gaiting has been implemented in certain instances [26, 15, 5] and also partially considered theoretically [14, 3, 10]. Not related to finger gaiting, however, perhaps the approach which most closely mirrors that of the subject of this paper is in [27] where notions of controllability and observability from "standard" control theory are applied to grasping. Additionally closely related is the work in [13] where the fundamental grasping constraint from [23] is slightly modified to include controlled relative velocities.

---

*This research is supported in part by NSF Grants No. IIS99-84107-CAREER and IIS99-10602-SGER.

2 Preliminaries

This presentation of background material assumes that the reader is familiar with concepts from grasping as well as nonlinear control from [16, 23].

2.1 Rolling Contact Kinematics

Most grasping motion planning or manipulation algorithms (such as [25]) or analysis (such as [27]) techniques which allow the fingers to roll relative to the object surface have been formulated in *contact coordinates.* In particular, the differential equations relating the evolution of contact to the relative velocities between a finger tip and object well known and are given by [22, 23]:

$$\begin{aligned} \dot{\alpha}_f &= M_f^{-1}\left(K_f+\tilde{K}_o\right)^{-1}\begin{bmatrix} -\omega_y \\ \omega_x \end{bmatrix} \\ \dot{\alpha}_o &= M_f^{-1}R_\psi\left(K_f+\tilde{K}_o\right)^{-1}\begin{bmatrix} -\omega_y \\ \omega_x \end{bmatrix} \\ \dot{\psi} &= \omega_z + T_f M_f \dot{\alpha}_f + T_o M_o \dot{\alpha}_o, \end{aligned} \tag{1}$$

where $\alpha_f = (u_f, v_f)$, $\alpha_o = (u_o, v_o) \in \mathbb{R}^2$ are the contact coordinates which parameterize the finger and object, respectively, the M's, T's and K's are the metric, torsion, and curvature forms, respectively, describing the geometry of the object or finger surfaces, denoted by subscript o or f, respectively, and the ω's are components of the relative angular velocities between frames affixed to the body and object at the point of contact.

The kinematic constraints which relate the relative velocities of the finger and object to joint velocities, which, coupled with Equation 1 provide a complete description of the manipulation dynamics. The coordinate frames used to describe the grasping manipulation are the standard frames from [23], and include the *palm frame*, P, a *station frame*, S_i, associated with each finger, a *finger frame*, F_i, associated with each finger and an *object frame*, O, used to describe the configuration of the object. Also, defined at every point on the surface of the object and finger tips are a family of Gauss frames, denoted by L_o and L_{f_i}. These frames are *fixed* with respect to the object and fingers, respectively. Also, define at the point of contact the Gauss frames C_o and C_{f_i} attached to the object and finger tips respectively which move with the point of contact.

Assuming that contact friction is sufficient to prevent slipping, then the directions in which forces can be applied are exactly the same components in which the relative velocity between L_o and L_{f_i} must be zero at the ith contact point, *i.e.*, $B^T V_{l_o l_f}^b = 0$, where $V_{l_o l_f}^b$ is the body velocity of frame L_f with respect to L_o and B is the *wrench basis* (see [23] for a complete explanation).

We consider a modified system where the wrench basis is appended with two additional columns that encode the fact that the relative rolling velocities (from Equation 1) are constrained to be a specified value. In particular, for point contact with friction,

$$B = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

where the fourth and fifth columns will constrain the x and y components of the relative angular velocity between the L frames on the object and finger. Note that for this contact model, only ω_z is unconstrained.

For the case where the relative angular velocities are going to be specified,

$$B^T V^b_{l_o l_f} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \omega_x \\ \omega_y \end{bmatrix} = \xi, \tag{2}$$

where ω_x and ω_y are determined (as illustrated subsequently) from Equation 1. A more useful expression will result from rewriting Equation 2 in terms of the velocity of the object and the joint velocity of the fingers, and a simple derivation results in

$$-B^T \mathrm{Ad}^{-1}_{g_{pl_f}} \mathrm{Ad}_{g_{po}} V^b_{po} + B^T \mathrm{Ad}^{-1}_{f_i l_f} J^b_{s_i f_i} \dot{\theta}_i = \xi, \tag{3}$$

where Ad is the adjoint transformation, V^b_{ab} is the body velocity of frame b relative to frame a and J^b_{ab} is the body Jacobian. Each is fully explained in [23].

2.2 Stratified Systems

One of the authors has previously considered nonholonomic motion planning and control for so-called stratified systems, which are systems that can switch among multiple contact states [11, 7, 8, 10, 6, 12]. Examples of stratified systems include the manipulation problem considered in this paper, legged locomotion and certain types of hybrid systems.

A simple example will provide an intuitive understanding of the geometry inherent in stratified systems. Consider the simplistic example two fingers intermittently engaging an object. The set of configurations corresponding to one of the robots engaging the object is a codimension one submanifold contained in the configuration space. The same is true when the other robot engages the object. Similarly, when both robots engage the object, the system is on a codimension two submanifold of the configuration space formed by the intersection of the single contact submanifolds. Each submanifold is referred to as a *stratum.* The structure of the configuration manifold for such a system is abstractly illustrated in Figure 1. Note that the equations of motion for the system will be *different* on each submanifold because the constraints on the system will be different on each submanifold.

By considering systems more general than the two cooperating robots in the example, a general definition of stratified configuration spaces can be developed. Let S_0 denote the system's entire configuration manifold and $S_i \subset S_0$ denote the codimension one submanifold of S_0 that corresponds to all configurations where only the ith robot engages the object. Denote, the intersection of S_i and S_j, by $S_{ij} = S_i \cap S_j$. The set S_{ij} physically corresponds to states where both the ith and jth robots engage the object. Further intersections can be similarly defined in a recursive fashion: $S_{ijk} = S_i \cap S_j \cap S_k = S_i \cap S_{jk}$, etc. The lowest-dimensional stratum (corresponding to all fingers in contact with the object for grasping problems) will be called the *bottom stratum.*

DEFINITION 2.1: (STRATIFIED CONFIGURATION MANIFOLD)
Let S_0 be a manifold, and n functions $\Phi_i : S_0 \mapsto \mathbb{R}$, $i = 1, \ldots n$ be such that the level sets $S_i = \Phi_i^{-1}(0) \subset S_0$ are regular submanifolds of S_0, for each i, and the intersection of any number of the level sets, $S_{i_1 i_2 \cdots i_m} = \Phi_{i_1}^{-1}(0) \cap \Phi_{i_2}^{-1}(0) \cap \cdots \cap \Phi_{i_m}^{-1}(0)$, $m \leq n$, is also a regular submanifold of S_o. Then S_0 and the functions Φ_n define a *stratified configuration space.*

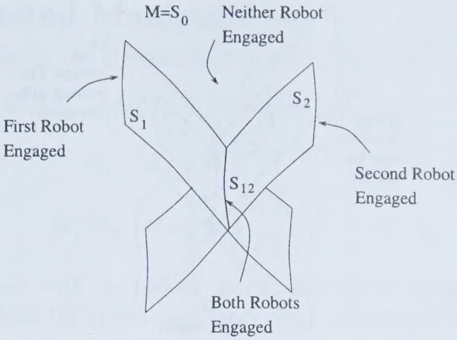


Figure 1. Configuration manifold structure for two cooperating robots.

2.3 Stratified Motion Planning

For smooth nonlinear systems, there are various motion planning techniques (piecewise constant inputs [17], steering with sinusoids [25, 24, 23, 36, 35], small amplitude inputs for mechanical system on Lie groups [19, 18, 2, 1], pushing, [20, 21], and others, [34, 33]). One of the authors has extended the method using piecewise constant inputs from [17] to the stratified case [12, 7] with application to legged robotic locomotion.

The basic approach is to consider the set of vector fields defined on the lowest-dimensional stratum (all fingers in contact) and to incorporate vector fields defined on higher strata by appropriately "projecting" them onto the bottom stratum. Then a series expansion (the *Chen–Flies* series) and the notion of the "extended system" (described subsequently) can be used in a straight–forward manner to construct control inputs which will steer the system to the final position with reference to a nominal trajectory.

For example, consider the simple cooperating robot configuration space as shown in Figure 1. Assume that on stratum S_{12}, (corresponding to both fingers in contact with the object) the vector field $g_{1,1}$ moves the system off of S_{12} and onto S_1, (finger 2 disengages the object) and correspondingly, $g_{2,1}$ moves the system off of S_{12} onto S_2 (finger 1 disengages the object). Also, consider the vector fields $g_{1,2}$ and $g_{2,2}$, defined on S_1 and S_2 respectively (corresponding to some motion of the system with fingers 2 and 1 not in contact with the object, respectively). Consider the following sequence of flows, starting from the point $x_0 \in S_{12}$

$$x_f = \underbrace{\phi^{t_6}_{-g_{2,1}}}_{S_{12}\leftarrow S_2} \circ \underbrace{\phi^{t_5}_{g_{2,2}}}_{\text{on } S_2} \circ \underbrace{\phi^{t_4}_{g_{2,1}}}_{S_2\leftarrow S_{12}} \circ \underbrace{\phi^{t_3}_{-g_{1,1}}}_{S_{12}\leftarrow S_1} \circ \underbrace{\phi^{t_2}_{g_{1,2}}}_{\text{on } S_1} \circ \underbrace{\phi^{t_1}_{g_{1,1}}}_{S_1\leftarrow S_{12}} (x_0), \tag{4}$$

as illustrated in Figure 2. The notation under each flow indicates what the flow is doing, *e.g.*, "$S_{12} \leftarrow S_1$" means that the flow takes the system from S_1 to S_{12} and "on S_1" means that the flow was entirely on S_1. In this sequence of flows, the system first moved off of the bottom stratum into S_1, flowed along the vector field $g_{1,2}$, flowed back onto the bottom stratum, off of the bottom stratum onto S_2, along vector field $g_{2,2}$ and back to the bottom stratum. In robotic finger gaiting, such a sequence of flows corresponds to the following sequence of motions:

1. finger 2 disengaging the object;

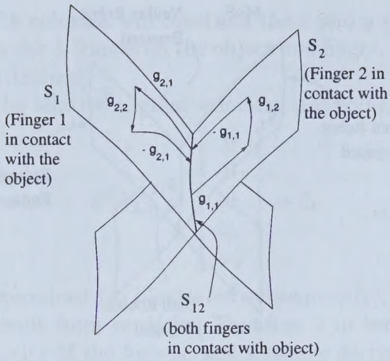

S_2
(Finger 2 in contact with the object)
$g_{2,1}$
S_1
$g_{2,2}$
$g_{1,2}$
(Finger 1 in contact with the object)
$-g_{1,1}$
$-g_{2,1}$
$g_{1,1}$
S_{12}
(both fingers in contact with object)

Figure 2. Sequence of flows.

2. some motion of the system with finger 1 in contact with the object and finger 2 not in contact with the object;
3. finger 2 engaging the object;
4. finger 1 disengaging the object;
5. some motion of the system with finger 2 in contact with the object and finger 1 not in contact with the object; and,
6. finger 1 engaging the object.

It is a basic result of differential geometry (the Campbell–Baker–Hausdorff formula), that if the Lie bracket between two vector fields is zero, then their flows commute. Thus, if

$$[g_{1,1}, g_{1,2}] = 0 \quad \text{and} \quad [g_{2,1}, g_{2,2}] = 0, \tag{5}$$

it is possible to reorder the above sequence of flows, by interchanging the flow along $g_{1,1}$ and $g_{1,2}$ and the flows along $g_{2,1}$ and $g_{2,2}$ as follows

$$x_f = \underbrace{\phi^{t_5}_{g_{2,2}} \circ \phi^{t_6}_{-g_{2,1}}}_{\text{interchanged}} \circ \phi^{t_4}_{g_{2,1}} \circ \underbrace{\phi^{t_2}_{g_{1,2}} \circ \phi^{t_3}_{-g_{1,1}}}_{\text{interchanged}} \circ \phi^{t_1}_{g_{1,1}}(x_0). \tag{6}$$

If $t_1 = t_3$ and $t_4 = t_6$, this reduces to

$$x_f = \underbrace{\phi^{t_4}_{g_{2,2}} \circ \phi^{t_2}_{g_{1,2}}}_{\text{on } S_{12}}(x_0). \tag{7}$$

Note that that $g_{1,2}$ and $g_{2,2}$ are vector fields in the equations of motion for the system on S_1 and S_2 respectively, (where each one of the fingers is not in contact), but *are not part of the equations of motion* on S_{12} when both fingers are in contact, but, for motion planning purposes can be considered as such because of the fact that Equation 6 results in the same net displacement as Equation 4, where the system switched between strata.

In the above example, the vector fields that took the system off of a substratum correspond in the grasping case to moving a finger out of contact (or back into contact) with the object. Due to the fact that the fingers are assumed to be holonomic, the Lie bracket decoupling expressed in Equation 5 will *always* be satisfied.

3 Underactuated Manipulation

The equations of motion for the grasping system are of the form

$$\begin{aligned} -B^T \mathrm{Ad}^{-1}_{g_{pl_f}} \mathrm{Ad}_{g_{po}} V^b_{po} + B^T \mathrm{Ad}^{-1}_{f_i l_f} J^b_{s_i f_i} \dot{\theta}_i &= \xi \\ M_f^{-1} \left(K_f + \tilde{K}_o\right)^{-1} \begin{bmatrix} -\omega_y \\ \omega_x \end{bmatrix} &= \dot{\alpha}_f \\ M_f^{-1} R_\psi \left(K_f + \tilde{K}_o\right)^{-1} \begin{bmatrix} -\omega_y \\ \omega_x \end{bmatrix} &= \dot{\alpha}_o, \end{aligned} \tag{8}$$

for each finger in contact with the object. For fingers out of contact with the object, there is no constraint from the object, and the contact coordinates evolve according to

$$\begin{aligned} \dot{\alpha}_f &= M_f^{-1} \left(K_f + \tilde{K}_o\right)^{-1} \left(\begin{bmatrix} -\omega_y \\ \omega_x \end{bmatrix} - \tilde{K}_o \begin{bmatrix} v_x \\ v_y \end{bmatrix} \right) \\ \dot{\alpha}_o &= M_0^{-1} R_\psi \left(K_f + \tilde{K}_o\right)^{-1} \left(\begin{bmatrix} -\omega_y \\ \omega_x \end{bmatrix} + K_f \begin{bmatrix} v_x \\ v_y \end{bmatrix} \right) \\ \dot{\psi} &= \omega_z + T_f M_f \dot{\alpha}_f + T_o M_o \dot{\alpha}_o, \end{aligned} \tag{9}$$

for each finger out of contact with the object and with $V^b_{po} = 0$. The states that we desire to control are the object velocity, V^b_{po} and the contact coordinates, α_f and α_o for each finger contact. Algebraically solving the above equations to separate them into an equation of the standard nonlinear control form:

$$\dot{x} = g_1(x)u_1 + \cdots + g_n(x)u_n, \tag{10}$$

and a set of state equations is theoretically straight-forward, but may be nearly intractable for complicated surface geometries.

A complete description of the motion planning algorithm from [17] and its extension to the stratified case in [7, 12, 9] is beyond the scope of this paper, so only a outline is provided here. Recall that the Lie bracket between two vector fields can be expressed by

$$[g_1, g_2](x) = \frac{\partial g_2(x)}{\partial x} g_1(x) - \frac{\partial g_1(x)}{\partial x} g_2(x),$$

and that the flow along the vector field corresponding to a Lie bracket motion can be approximated by

$$\phi^t_{[g_1,g_2]}(x_0) \approx \phi^{\sqrt{t}}_{-g_2} \circ \phi^{\sqrt{t}}_{-g_1} \circ \phi^{\sqrt{t}}_{g_2} \circ \phi^{g_1}_{\sqrt{t}}(x_0), \tag{11}$$

where ϕ^t_g represents the flow of the system along the vector field g for time t, *i.e.*, the control corresponding to vector field g is turned on for time t. Although the construction and formalism is substantial, the basic idea in [17] is to decompose a desired motion (called the nominal trajectory) into multiple subtrajectories along various vector fields that provide a basis for space of vector fields. If the system is underactuated, some of these elements will be Lie brackets, and flowing in Lie bracket directions will need to be approximated in a manner expressed in Equation 11.

In particular, related to the original system (Equations 8 and 9) is *formal differential extended system*

$$\dot{S}(t) = S(t)\left(B_1 v_1 + \cdots + B_p v_p\right) \tag{12}$$

where the B_i's belong to a noncommutative formal Lie algebra and are related to the original vector fields in Equation 10 and their Lie brackets. The extended system is picked so that the vector fields corresponding to the B_i's are full rank at every point along the desired trajectory. Therefore, if the original system is underactuated, the extended system will contain Lie brackets directions along which the system cannot directly flow. Additionally, all flows of the original system can be represented (formally) by

$$S(t) = e^{h_p(t)B_p} e^{h_{p-1}(t)B_{p-1}} \cdots e^{h_1(t)B_1}, \tag{13}$$

where the h_i are called the *backward Philip Hall coordinates* and, because of the formal representation, the exponentials can be expanded in the "standard" series expansion for exponentials. Differentiating Equation 13 with respect to time and equating the resulting coefficients of the B_i's with the coefficients of the B_i's in Equation 12 yields differential equations for that can be solved to determine the backward Philip Hall coordinates. Once the Philip Hall coordinates are computed, it is straight-forward to construct piecewise constant control inputs for the *original system* (Equation 9) to approximate the total flow of the system along γ, as was illustrated by Equation 11.. The method works exactly for *nilpotent systems* (nilpotency is a property of the Lie algebra containing the g_i). For the general, non-nilpotent case, the method works approximately and [17] derives explicit bounds on the resulting error.

Now, in the stratified grasping case, since the Lie bracket decoupling expressed in Equation 5 is always satisfied, motions when each of the fingers are out of contact with the object can be considered as part of the collection of vector fields that can be used for motion planning. Accordingly, we can define the *extended stratified system.*

DEFINITION 3.1: (EXTENDED STRATIFIED SYSTEM)
The *extended stratified system* on the bottom strata, S_B, is the driftless system comprised of the vector fields on the bottom strata, chosen vector fields from the higher strata, and Lie brackets of vector fields from S_B and higher strata, *i.e.*, it is a system taking the form:

$$\dot{x} = b_1(x)v_1 + \cdots b_m(x)v_m + \underbrace{b_{m+1}v_{m+1} \cdots + b_n v_n}_{\text{from higher strata}} + \underbrace{b_{n+1}v_{n+1} + \cdots + b_p v_p}_{\text{any Lie brackets}}, \tag{14}$$

where the $\{b_1, \ldots, b_p\}$ span $T_x S_0$, the inputs $v_1, \ldots, v_n$ are real, and the inputs $v_{n+1}, \ldots, v_p$ are fictitious.

Specifically, the algorithm to generate trajectories that move the system from initial configuration p to final configuration q is as follows.

1. Construct the extended stratified system, Equation (14), on the bottom strata, S_B.
2. Find a nominal trajectory, $\gamma(t)$, that connects p and q. Given $\gamma(t)$, solve

   $$\dot{\gamma}(t) = b_1(x)v_1 + \cdots + b_p(x)v_p,$$

   for the fictitious inputs, v_i.
3. Solve the stratified extended system for the fictitious control inputs, *i.e.*, solve for the backward Philip Hall coordinates by solving the differential equations derived from from Equations 12 and 13.
4. For each path segment in each strata, compute the actual controls that steer the system along $\gamma(t)$.

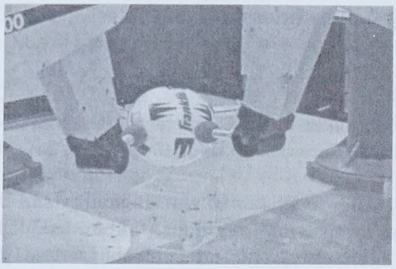
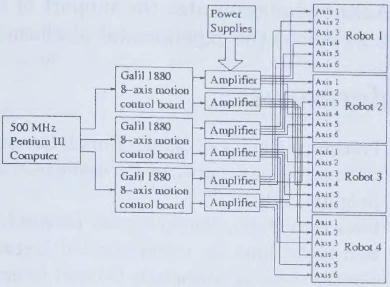


Figure 3. Experimental stratified manipulation.



Power Supplies
Galil 1880 8–axis motion control board
Galil 1880 8–axis motion control board
Galil 1880 8–axis motion control board
500 MHz Pentium III Computer
Amplifier
Amplifier
Amplifier
Amplifier
Amplifier
Amplifier
Axis 1
Axis 2
Axis 3
Axis 4
Axis 5
Axis 6
Robot 1
Robot 2
Robot 3
Robot 4

Figure 4. Schematic of control system.

5. Flow along each first order vector field, and approximate higher order vector fields as illustrated in Equation 11. In general, it will be necessary to switch strata between some of these flows.

4 Experimental Validation

The above results have also been verified and demonstrated experimentally. The experimental platform consists of four standard Puma 560 robots mounted on a common platform. All of the robots are controlled by a central 500 MHz Pentium III computer via Galil 1880 8-axis motion control boards operating six amplifies, each controlling four axes each. A close-up of the four robots with spherical finger tips engaging a spherical football is illustrated in Figure 3 and the complete system is schematically illustrated in Figure 4. Experiments are carried out with spherical and "egg-shaped" objects as spherical and flat finger tips with only four of the robot axes actuated. Any arbitrary axis of rotation can be specified and the ensuing motion is extremely robust and precise where the robots are able to completely rotate the ball several times with many instances of the fingers coming in and out of contact. Given that the algorithm is open loop, such a level of precision and robustness was surprising even to the authors. Movies of sample experimental results are available via the world wide web at `http://controls.ame.nd.edu/manip/`.

5 Conclusions

This paper presented an outline of a robotic manipulation planning technique to implement so-called "finger gaiting." It is based on nonholonomic motion planning techniques which have been extended by the authors to a class of discontinuous problems which includes robotic grasping. It is formulated in a level of mathematical generality so that the algorithm can accommodate any smooth object and finger tip parameterizations and arbitrary controllable kinematics of the manipulators. Simulation as well as experimental results were presented. Not presented, but a direct result from the authors' previous efforts is a proof that force closure can be maintained throughout the manipulation process. Interested readers are referred to [12, 7] for an applicable proof.

Future work includes adopting a vision-based robotic control method to "close the loop" to further enhance robustness and precision.. Additionally, a current requirement of the algorithm is that the object and finger tip be smoothly parameterized. Work to extend the algorithm to non-smooth objects will be the subject of a future publication.

Acknowledgements: the support of the National Science Foundation for both authors as well as the experimental platform is gratefully acknowledged.

References

[1] Francesco Bullo and Naomi Ehrich Leonard. Motion control for underactuated mechanical systems on lie groups. In *Proceedings of the 1997 European Control Conference*, Brussels, Belgium, 1997.

[2] Francesco Bullo, Naomi Ehrich Leonard, and Andrew D. Lewis. Controllability and motion algorithms for underactuated lagrangian systems on lie groups. Submitted, IEEE Transactions on Automatic Control Issue on Mechanics, Nonlinear Control Systems, 1999.

[3] I-M Chen and J.W. Burdick. A qualitative test for n–finger force–closure grasps on planar objects with applications to manipulation and finger gaits. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 814–820, 1993.

[4] Ronald S. Fearing and T.O. Binford. Using a cylindrical tactile sensor for determining curvature. *IEEE Transactions on Robotics and Automation*, 7(6):806–817, 1991.

[5] R.S. Fearing. Implementing a force strategy for object reorientation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 96–102, 1986.

[6] Bill Goodwine. Stratified motion planning with application to robotic finger gaiting. Proceedings of the 10th World Congress on the Theory of Machines and Mechanisms, 1999.

[7] Bill Goodwine and Joel Burdick. Trajectory generation for kinematic legged robots. In *IEEE International Conference on Robotics and Automation*, pages 2689–2696, Albuquerque, NM, 1997.

[8] Bill Goodwine and Joel Burdick. Gait controllability for legged robots. In *Proceedings of the 1998 International Conference on Robotics and Automation*, Leuven, Belgium, 1998. IEEE.

[9] Bill Goodwine and Joel Burdick. A general method for motion planning for quasi–static legged robotic locomotion and figer gaiting. Submitted to the IEEE Transactions on Robotics and Automation, available ectronically at `http://controls.ame.nd.edu/~bill/papers.html`, 1998.

[10] Bill Goodwine and Joel Burdick. Stratified motion planning with application to robotic finger gaiting. To appear in the Proceedings of the 1999 IFAC World Congress, 1999.

[11] Bill Goodwine and Joel Burdick. Controllability of kinematic systems on stratified configuration spaces. To appear in IEEE Transactions on Automatic Control, June, 2001, available electronically at `http://controls.ame.nd.edu/~bill/papers.html`, 2001.

[12] J. William Goodwine. *Control of Stratified Systems with Robotic Applications.* PhD thesis, California Institute of Technology, 1998.

[13] L. Han, Y.S. Guan, Z.X. Li, Q. Shi, and J.C. Trinkle. Dextrous manipulation with rolling contacts. In *Proceedings of the 1997 IEEE International Conference on Robotics and Automation*, pages 992–997. IEEE, 1997.

[14] J. Hong and G. Lafferriere. Fine manipulation with multifinger hands. In *IEEE International Conference on Robotics and Automation*, pages 1568–1573, Cincinnati, OH, May 1990.

[15] Maw Kae Hor. *Control and Task Planning of the Four Finger Manipulator.* PhD thesis, NYU, 1987.

[16] Alberto Isidori. *Nonlinear Control Systems.* Springer–Verlag, second edition, 1989.

[17] G. Lafferriere and Hector J. Sussmann. A differential geometric approach to motion planning. In X. Li and J. F. Canny, editors, *Nonholonomic Motion Planning*, pages 235–270. Kluwer, 1993.

[18] N.E. Leonard. Periodic forcing, dynamics and control of underactuated spacecraft and underwater vehicles. In *Proceedings of the 34th IEEE Conference on Decision and Control*,

1995.

[19] N.E. Leonard and P.S. Krishnaprasad. Motion control of drift–free, left–invariant systems on lie groups. *IEEE Transactions on Automatic Control*, 40(9):1539–1554, September 1995.

[20] K. M. Lynch. *Nonprehensile Robotic Manipulation: Controllability and Planning.* PhD thesis, Carnegie Mellon University, 1996.

[21] K.M. Lynch and M.T. Mason. Stable pushing: Mechanics, controllabilty, and planning. *International Journal of Robotics Research*, 15(6):533–556, December 1996.

[22] D. J. Montana. The kinematics of contact and grasp. *The International Journal of Robotics Research*, 7(3):17–25, 1988.

[23] Richard M. Murray, Zexiang Li, and S. Shankar Sastry. *A Mathematical Introduction to Robotic Manipulation.* CRC Press, Inc., 1994.

[24] Richard M. Murray and S. Shankar Sastry. Grasping and manipulation using multifingered robot hands. In *Robotics*, volume 41 of *Proceedings of Symposia in Applied Mathematics*, pages 91–128. American Mathematical Society, Providence, Rhode Island, 1990.

[25] R.M. Murray and S.S. Sastry. Nonholonomic motion planning: Steering using sinusoids. *IEEE Transactions Automatic Control*, 38:700–716, 1993.

[26] T. Okada. Object handling system for manual industry. *IEEE Transactions on Systems, Man and Cybernetics*, 9(2):79–89, 1979.

[27] Domenico Prattichizzo and Antonio Bicchi. Dynamic analysis of mobility and graspability of general manipulation systems. *IEEE Transactions on Robotics and Automation*, 14(2):241–258, April 1998.

[28] E. Rimon and J.W. Burdick. Configuration space analysis of bodies in contact – i. *Mechanism and Machine Theory*, 30(6):897–912, August 1995.

[29] E. Rimon and J.W. Burdick. Configuration space analysis of bodies in contact – ii. *Mechanism and Machine Theory*, 30(6):913–928, August 1995.

[30] K. Salisbury, D. Brock, T. Massie, N. Swarup, and C. Zilles. Haptic rendering: Programming touch interation with virtual objects. In *Proceedings of the Symposium on Interactive 3D Graphics*, pages 123–130, 1995.

[31] Kenneth Salisbury and Christopher Tarr. Haptic rendering of surfaces defined by implicit functions. In *Proceedings of the 1997 ASME International Mechanical Engineering Congress and Exposition*, pages 61–67, 1997.

[32] K.B. Shimoga. Robot grasp synthesis algorithms: A survey. *The International Journal of Robotics Research*, 15(3):230–266, 1996.

[33] H.J. Sussmann and W. Liu. Limits of highly oscillatory controls and the approximation of general paths by admissible trajectories. In *Proceedings of the 30th IEEE Conf. Decision and Control*, pages 437–442, Brighton, UK, 1991. IEEE.

[34] H.J. Sussmann and W. Liu. Motion planning and approximate tracking for controllable systems without drift. In *Proc. 25th Annual Conf. on Inf. Sciences and Systems*, pages 547–551, Johns Hopkins University, 1991.

[35] D. Tilbury, R. Murray, and S. Sastry. Trajectory generation for the n-trailer problem using foursat normal form. UC Berkeley ERL Memo Number M93/12., 1993.

[36] D. Tilbury, R. M. Murray, and S. Sastry. Trajectory generation for the n-trailer problem using goursat normal form. In *Proceedings of the 32nd IEEE Conference on Decision and Conrtrol*, San Antonio, TX, 1993. IEEE.

[37] J.C. Trinkle and R.P. Paul. Planning for dexterous manipulation with sliding contacts. *The International Journal of Robotics Research*, 9(3):24–48, 1990.

Real-Time Motion Planning for Autonomous Vehicles

E. Frazzoli, M. A. Dahleh, E. Feron
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology
email:`{frazzoli,dahleh,feron}@mit.edu`

Abstract

In this paper we will present a randomized algorithm for motion planning, based on the probabilistic roadmap approach, which directly takes into account in a very efficient manner the dynamics of the vehicle. This is accomplished through the introduction of a fundamental concept in control theory: the construction of the probabilistic roadmap will be performed through the sequencing of Lyapunov functions. The main advantage of this approach is the capability of addressing directly the dynamics constraints on the vehicle's motion, while at the same time providing a consistent decoupling between low-level control and motion planning. Simulation examples, involving autonomous spacecraft and a small autonomous helicopter, will be presented and discussed.

1 Introduction

One of the basic problems which have to be faced by autonomous vehicles or moving robots is the generation and the execution of a motion plan, that enables the robot to move to a desired location, while avoiding collisions with obstacles, or other undesirable behaviors. When dealing with realistic situations, the robot will have to generate and execute this plan in real-time, in an environment with a complex topology, and with dynamically changing and uncertain components. Moreover, we are often interested in devising motion planning strategies for agile autonomous vehicles, operating in an environment such that the exploitation of their dynamics and maneuvering capabilities plays a crucial role in achieving the mission goals. In such cases motion planning strategies should take into full account the dynamics of the vehicle. In other words, it is desirable that the output from the motion planning be compatible with the system dynamics, i.e. a feasible trajectory.

A review of the current status of motion planning research appeared in [1], and a comprehensive treatment of the problem can be found in [2, 3, 4]. The computational complexity of a deterministic, complete algorithm, even in the simplest cases of motion planning, has been proven to be such to require exponential time in the number of the degrees of freedom of the vehicle. In order to circumvent the comlutational complexity of complete algorithms, a new class of motion planning algorithms, known as probabilistic roadmap planners, has been recently introduced in [5]. This class of algorithms is particularly interesting since, by relaxing the completeness requirements to probabilistic completeness (i.e. completeness in expectation), one can achieve computational tractability, while maintaining formal guarantees on the behavior of the algorithm. In this paper, an extension of the available algorithms is presented, which is able to deal effectively with the system's dynamics, in an environment characterized by moving obstacles.

2 Planning framework

In the following we will present different modeling approaches for describing the system dynamics. While the continuous state-space formulation is the most traditional and widely used representation of a vehicle's dynamics, recent advances in hybrid systems theory suggest an alternative formulation, which presents significant computational advantages, especially for aerospace vehicles with complex, high-dimensional and fast dynamics.

2.1 State space formulation

The usual representation of the dynamics of an autonomous vehicle or robot is based on a state-space formulation. The dynamics of the system are described by an ordinary differential equation of the form:

$$\frac{dx}{dt} = f(x, u) \tag{1}$$

where $x \in \mathcal{X}$ is the state, belonging to an n-dimensional manifold $\mathcal{X}$ (the *state space*), and u represents the control input signals, taking values in the set $\mathcal{U} \subseteq \mathbb{R}^m$. The function $f : \mathcal{X} \times \mathcal{U} \to T\mathcal{X}$ is assumed to be locally Lipschitz in its arguments. Note that the above formulation can represent both nonholonomic and dynamic constraints[6]. In some cases, we also have to consider additional inequality constraints on the state variables $F_i(x) < 0$, to ensure safe operation of the vehicle (e.g. flight envelope protection).

Assume that the state space $\mathcal{X}$ can be decomposed, at least locally, in the product $\mathcal{C} \times \mathcal{V}$. We can regard $\mathcal{C}$ as the *configuration space* of the vehicle, while $\mathcal{V}$ encodes the vehicle's velocity and higher order derivatives of the configuration variables.

A significant reduction in the complexity of the motion planning problem can be achieved under the assumption that the system dynamics are *invariant* with respect to translations (or, in general, group operations) on $\mathcal{C}$. One of the main consequence of the invariance (or symmetry) is the existence of relative equilibria, or trim trajectories[7]. While the motion planning problem can be posed in terms of relative equilibria, for the sake of simplicity, in this paper we will limit the analysis to problems in which the desired destination is an *equilibrium point*. We can define equilibrium points for the system (1) as the points $(\bar{x}, \bar{u})$ for which $f(\bar{x}, \bar{u}) = 0$. Since the system dynamics are invariant with respect to translations on $\mathcal{C}$, a family of equilibrium states can be expressed by a point in $\mathcal{C} \times \{\bar{v}\}$, where $\bar{v} \in \mathcal{V}$ is a constant (e.g. the zero vector).

2.2 Hybrid formulation

While the state space formulation is probably the most commonly used representation of a vehicle's dynamics, it could not be the most appropriate choice in all cases. In particular, the state space of non-trivial systems is typically very large, and the "curse of dimensionality" makes the solution of motion planning problems in such large-dimension spaces computationally intractable.

An alternative approach is represented by the formulation of the system dynamics, and hence of the motion planning problem, in what can be regarded as the *maneuver space* of the vehicle[8, 9]. Such an approach entails the definition of a hybrid control architecture (a *hybrid automaton*), based on quantization of the system dynamics into a library of feasible "trajectory primitives", or behaviors. We will not discuss in detail the architecture proposed in the references, it will suffice to mention its main features.

First of all, we have to identify the trajectory primitives that will be included in the library. As mentioned in the previous section, from invariance to translation in $\mathcal{C}$, it follows that there exists a particular class of trajectories, known as *relative equilibria*. An appropriately chosen set of relative equilibria constitutes the first class of primitives. The elements of the second class of primitives, that we can call *maneuvers*, are defined as feasible, finite time transitions between trim trajectories.

We can graphically depict the hybrid automaton as a directed graph, where the nodes represent the trim trajectories, and the edges represent the maneuvers. If we restrict the evolution of the vehicle to follow trajectories generated by the automaton, the complete state will be described by the index of the current trajectory primitive, the current time, and the time and location on $\mathcal{C}$ at the primitive inception time. Explicit conditions can be given to ensure reachability of the resulting dynamical system, as well as consistency or robustness to uncertainties in the environment and in the plant[9].

Note that, while maneuvers have a definite time duration, trim trajectories can be followed indefinitely. As a consequence, the decision variables, or control inputs, acting on the hybrid automaton, consist of the amount of time the vehicle must remain in the trim trajectory (*coasting time*), and in which maneuver it has to execute next (*jump destination*).

The hybrid automaton can be seen as a new *modeling tool*, in which the continuous dynamics ODE (1) is replaced by the transition rules on the directed graph representing the automaton, and by the associated hybrid system evolution[8, 9]. Using this representation, the assumption of invariance to translations in $\mathcal{C}$ is enough to ensure that the hybrid automaton encodes all the relevant information about dynamics and the flight envelope constraints of *any* vehicle in a space of small dimension, that is $\mathcal{M} = \mathcal{C} \times Q_T$. For most vehicles, the dimension of $\mathcal{C}$ is at most 4, and Q_T is a finite set of trim trajectory indices. In the rest of the paper, to simplify the notation, we will just use the letter $\mathcal{X}$ to indicate the "state space", with the understanding that this can be regarded as either the proper, continuous state space or the hybrid maneuver space $\mathcal{M}$.

2.3 Environment characterization

We will consider an environment which presents both fixed and moving obstacles, and we will assume that the motion of the obstacles (or conservative estimates thereof) is *known in advance*. Since the environment is time-varying, collisions have to be checked on (space × time) pairs $(x, t) \in \mathcal{X} \times \mathbb{R}$.

Given an initial condition (x_0, t_0), a pair (x_f, t_f) is said *reachable* if it is possible to find a control function $\hat{u} : [t_0, t_f] \to \mathcal{U}$, such that the ensuing trajectory of the system, from the above initial conditions is feasible, collision free, and terminates at the required final condition.

2.4 Problem formulation

The motion planning problem we are interested in can now be stated as: given an initial state $x_0 \in \mathcal{X}$, at time t_0, and a goal *equilibrium* configuration $x_f \in \mathcal{C} \times \{\bar{v}\}$, find a control input $v : [t_0, t_f] \to \mathcal{U}$, such that the resulting trajectory χ is feasible, collision free, and satisfies the boundary conditions $\chi(t_0) = x_0$, $\chi(t_f) = x_f$. A motion planning algorithm is said *complete* if it returns a feasible solution if there exists a time t_f such that (x_f, t_f) is reachable from (x_0, t_0), and returns failure otherwise. While the usual formulation of the motion planning problem is concerned only with finding a feasible trajectory, in many engineering applications we are also interested in finding a trajectory minimizing some cost, often expressed as a functional of the form $J := \int_{t_0}^{t_f} g(\chi(t), v(t))dt + \Gamma(x(t_f), t_f)$.

3 Motion planning algorithm

Before addressing the problem of motion planning in the presence of obstacles, we will discuss the solution to the problem of optimal motion planning in an obstacle-free environment. In the following, we will try to make it clear how the solution to this problem will be instrumental in devising (sub)optimal trajectories when obstacles are included.

3.1 Planning in an obstacle-free environment

Consider the problem of finding the control input that, given a target *equilibrium point* x_{eq}, minimizes the total cost:

$$J(x_0, x_{eq}) = \int_{t_0}^{t_f} g(x, x_{eq}, u)dt + \Gamma(x(t_f), x_{eq}, t_f) \tag{2}$$

for some initial conditions x_0, under the dynamics constraints (and possibly state and control constraints). Assume that we have solved the above optimal control problem, and that we know the optimal cost function $J^*(x, x_{eq})$, for all $x \in \mathcal{X}$, and all equilibrium points $x_{eq} \in \mathcal{C} \times \{\bar{v}\}$: then it is relatively easy to formulate an optimal control policy $\pi : \mathcal{X} \times \mathcal{C} \to \mathcal{U}$, as a feedback policy that returns at each time instant the control input that minimizes the cost-to-go to the target.

The solution to an optimal control problem in the free space thus provides us with a control policy π that ensures that the system is driven towards an equilibrium point, effectively parameterized by configurations in $\mathcal{C}$.

3.2 Motion Planning in the presence of obstacles

The motion planning algorithm in the presence of obstacles that we will present in the following, first introduced in [10], is based on the determination of a time-parametrized sequence of "attraction points" $x_{eq}(t)$ that effectively steers the system to the desired configuration while avoiding obstacles. Since the actual control inputs can be computed from the knowledge of the optimal control policy $\pi(\cdot, x_{eq})$, this means that the low-level control layer (the layer actually interacting with the vehicle) and the high-level, guidance layer are effectively decoupled, while at the same time ensuring full consistency between the two levels. An analysis of the probabilistic completeness of the algorithm, together with performance bounds, is available in [10]

The basic algorithm can be outlined as follows: Starting with a node representing the initial condition (at some point $\tau > 0$ seconds in the future), we build a tree by iteratively adding new "milestones", which are connected to the tree by a feasible, and provably safe, trajectory segment. At the tree nodes we store all the information concerning each milestone. More specifically, the data stored at the tree nodes include the propagated state of the vehicle (i.e. state $x \in \mathcal{X}$, time $t \in \mathbb{R}$), plus information on the "quality" of the motion plan, such as the cumulative cost and estimates on the cost-to-go. We also include a counter of the total number of milestones in the children trees. The (state × time) couple is initialized through propagation of the system dynamics, and the cumulative cost is updated, according to eq.(2). The lower bound on the cost-to-go coincides with the value of the cost function

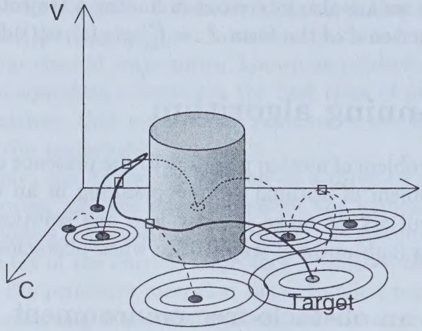


Figure 1: Example of generated roadmap (projected on $\mathcal{X}$). Primary milestones are marked as circles, and secondary milestones as squares

$J^*(x, x_f)$, while the upper bound on the cost to go is initialized to $+\infty$, meaning that a feasible path from the particular node has not been found yet.

The first step in the generation of a new milestone is the selection of a random equilibrium state $x_{rand} \in \mathcal{C} \times \{\bar{v}\}$; the control policy $\pi(\cdot, x_{rand})$ is applied to a randomly chosen node of the current tree, and if the ensuing propagated trajectory is feasible (and the resulting final state is τ-safe, see below), x_{rand} is added to the tree. The milestones generated according to the above procedure can be regarded as *primary* milestones.

In order to increase the performance of the resulting trajectory, we will introduce the following step: at random points along the trajectory to the new randomly generated point, we add $n_s \geq 1$ additional, *secondary*, milestones. Secondary milestones are likely to be at points in the state space that are "far" from equilibria. From secondary milestones we can apply the control policy to the destination $\pi(\cdot, x_f)$: if the resulting trajectory is feasible, we have solved the feasibility problem.

In this case, we climb the tree back towards root, updating the estimates on the upper bound on the cost-to-go. Both in the case in which a feasible trajectory is found, and in the case in which a collision is detected, the secondary milestones are added to the tree, and can be selected as the starting point for later iterations. Note that each secondary milestone, by construction, will have a primary milestone in a child subtree (see Fig. 1).

In order to guard ourselves from dead-ends due to finite computation times, we also require that primary milestones be checked for τ-safety. We say that a milestone $(x, t) \in \mathcal{C} \times \{\bar{v}\} \times \mathbb{R}$ is *τ-safe* if $(x, \tilde{t}) \in \mathcal{F}$ for all $\tilde{t} \in [t, t + \tau]$. The time horizon τ should be chosen as a time interval in which at least one new τ-safe milestone can be computed with high probability. Note that if it is possible to check safety on an infinite horizon (for example in the case of a static environment), the τ-safety requirement translates into a hard safety guarantee on the computed solutions.

The time available for computation is bounded by either τ, or by the duration of the current trajectory segment. When the time is up, we have to select a new tree from the children of the current root. If there are none, we know that every primary milestone will be τ-safe, and hence we will have τ seconds for computing a new tree (secondary milestones will *always* have at least one child). If there are children, we have two possibilities. In the most favorable case, at least one of the children will lead to the destination through an already computed feasible solution. If there is more than one such feasible solution, we select the one with the least upper bound on the cost-to-go. On the other hand, it can be the case (especially during the first iterations of the search), that no feasible solutions have been computed yet. In this case we randomly select the child to descend according to a distribution weighted on the total number of milestones in each tree. The selected tree will be likely to cover a bigger portion of the reachable set.

3.3 Tree pruning

In the previous section we have mentioned how at each node we can store an upper and a lower bound on the cost-to-go. The lower bound coincides with the optimal cost-to-go in the obstacle-free case, whereas the upper bound is either a finite, positive number if a feasible trajectory has been found and is among the children of the current node, or $+\infty$ if not. The presence of upper and lower bounds on the cost-to-go makes it possible to apply ideas and techniques from branch-and-bound algorithms: in our case, we can easily prune the tree from provably "bad candidates".

Every time a new feasible solution is found, we can update the upper bounds on the cost-to-go by climbing the tree backwards towards root. While performing this operation, we can also look at all the children of the node being updated. If the lower bound on the total cost-to-go for such a children is higher than the upper bound on the cost-to-go for the current node, we can remove the corresponding subtree.

The end result of such a process is the removal from the trajectory tree of all the provably bad candidates for the "optimal" solution. The trajectory tree, following this pruning process, will contain a smaller number of nodes, thus improving the overall computational efficiency. However, it must be kept in mind that tree pruning can only be carried out once a feasible solution has been found, and is of no help before that happens.

3.4 Time-proximity heuristic

Some heuristic criteria can be introduced to improve the performance of the algorithm in finding a feasible solution. Even though the basic algorithm presented in this paper can be proven to have an exponential convergence rate [10, 11], in practice it suffers from a major drawback: since the selection of milestones, and of the tree nodes to be expanded, is carried out completely at random, the resulting trajectory tree lacks any "sense of purpose" in the exploration. On the other hands, other algorithms (e.g. the RRT in [12]) select the node to be expanded as the node that is closest to the randomly generated milestone. This results in a rapid exploration of the reachable space, since the RRT is biased towards the reduction of the largest Voronoi diagram associated with the trajectory tree. Unfortunately, this technique could not work for a dynamic environment, since it is based only on a notion of distance in the configuration space. In the following, we will use some ideas from the RRT algorithm, which apply to the dynamic environment case.

First of all, we can introduce a notion of "time proximity" as follows. Let's consider the function $T^*_\pi : \mathcal{X} \times \mathcal{C} \to \mathbb{R}_+$, where $T^*_\pi(x, x_{eq})$ returns the time necessary to reach the equilibrium point x_{eq} from the initial condition x, using the policy π in an obstacle-free environment. Notice that in a minimum-time formulation of the optimal control problem, this is the value function of the obstacle-free case. Define now a target milestone as a couple (x_{rand}, t_{rand}). Given a node in the tree (x, t), we can define the *time distance* δ_t between the node and the milestone as:

$$\delta_t := \begin{cases} t_{rand} - t & \text{if } t_{rand} - t > T^*_\pi(x, x_{rand}) \\ +\infty & \text{otherwise} \end{cases} \tag{3}$$

In other words, the distance is finite and equal to the time difference between the two (state,time) couples if and only if the destination is reachable from the starting point in the obstacle-free case.

Now we can introduce the following heuristic: given a randomly generated milestone (x_{rand}, t_{rand}) choose the node to be expanded as the node that is closest to the milestone, in the sense defined by δ_t. In order to ensure that δ_t is finite for at least one node of the tree, we can choose t_{rand} as follows:

$$t_{rand} = T^*_\pi(x_{root}, x_{rand}) + \tilde{\tau} \tag{4}$$

where $\tilde{\tau} > 0$ is a randomly generated delay.

Even though a formal analysis has not been performed yet, the application of this heuristic method for choosing nodes to expand resulted in a significant reduction of the time required to find a feasible trajectory.

4 Application Examples

In the next two sections we will present two examples which show the power and the flexibility of the proposed algorithm. We will consider first a dynamic system which evolves on a Lie group (namely, the rotation group), and then a system endowed with a control architecture patterned after the hybrid control structure described in section 2.2. All algorithms have been implemented in C++ on a Pentium II 300 MHz machine. Computer animations of the simulation examples are available from the authors.

4.1 Spacecraft attitude motion

A key enabling technology that could lead to greater spacecraft autonomy is the capability to autonomously and optimally slew the spacecraft between attitudes while operating under a number of celestial and dynamic constraints. For example, when planning a spacecraft slew maneuver, great care has to be taken in order to protect sensitive science or navigation sensors (e.g. telescopes, star trackers) from direct exposure to the Sun and other bright bodies. The task of finding an attitude trajectory that meets all the constraints is a formidable one: in fact, attitude maneuver planning has traditionally been performed on the ground.

The motion planning algorithm outlined in this paper can be applied directly to the problem of spacecraft slewing. In our algorithm the connection between the system dynamics and the motion planning algorithm is provided by the control policy π defined in section (3.1). We can identify the configuration space $\mathcal{C}$ as the rotation group $SO(3)$; in the case in which we can model the spacecraft as a rigid body, the "velocity space" $\mathcal{V}$ can be identified as the associated Lie algebra $\mathfrak{so}(3)$ (angular velocities in body axes).

Several approaches have been proposed for designing attitude control policies. In the examples presented in the following, we used the coordinate-independent control laws derived in [13], which provide a large set of attraction for equilibrium points. These control laws are appropriate for the case in which the spacecraft is equipped with momentum wheels, providing independent torques on three axes.

The formulation of obstacles, and the "collision checking" function used to ensure feasibility of a planned trajectory, can be easily extended to include a wide range of pointwise in time constraint checking. Moreover, the fact that we are building a tree rooted at some initial condition, allows easy implementation of integral constraints, by simple bookkeeping at the tree nodes and edges.

In the following example, the case of a deep-space spacecraft, cruising in the space between the orbits of Mars and Jupiter, is considered. The spacecraft carries a telescope on the $+Z$ axis, and a star tracker on the $-Z$ axis. The telescope boresight cannot be less than 30 degrees off the Sun, whereas the star tracker boresight has to stay at least 40 degrees away from any bright object, including the Sun, as well as the Earth, Mars, and Jupiter. Limits are imposed on angular velocity, control torque, and total stored momentum in the reaction weels. An attitude slew is commanded, for which a direct path would result in a violation of the constraints. A feasible solution was found by the proposed algorithm in about 1.1 seconds on average (standard deviation 0.6 seconds). An example of a planned trajectory is depicted in figure 4.1.

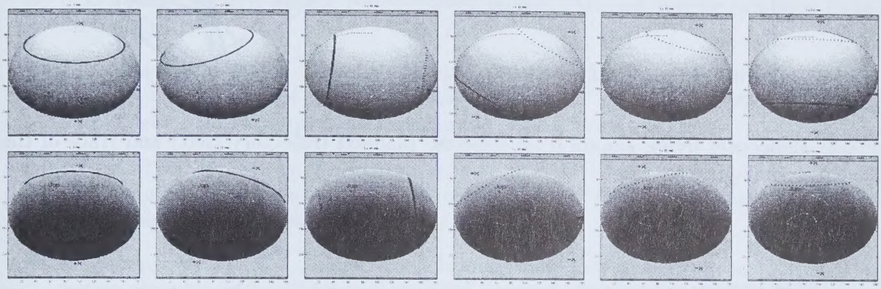


Figure 2: Attitude motion planning example. The top and bottom rows show the same maneuver, from view points on the opposite sides of the sphere

4.2 Small autonomous helicopter

One of the prime examples of systems for which our algorithm is applicable is a small autonomous helicopter. Radio-controlled helicopters are capable of remarkably agile and aggressive maneuvers, which are impossible to perform autonomously using traditional control techniques, especially when the on-line solution of the motion planning problem is required. Among the reasons for that we can state that helicopters are essentially unstable systems, with a very high bandwidth, and their dynamics change considerably throughout the flight envelope.

In this section, we will present simulation results for a test case involving a small autonomous helicopter. The simulations are carried out on a fully non-linear helicopter simulation, based on a widely used minimum-complexity model[14]. The motion planning algorithms operating on the hybrid automaton structure[8, 9] are complemented by the control law presented in [15] to ensure tracking of the reference trajectory.

The randomized path planning has been tested in several examples, including cases with both fixed and moving obstacles, and in general proved to be very fast and reliable. The cost function used in all the examples is the total time needed to go to the destination (we are solving a minimum-time problem). The first example involves navigating the helicopter through a set of obstacles represented as spheres in the configuration space. This example was very easily handled by the proposed planner: a feasible solution was found on average in 136 ms (standard deviation 59 ms).

In the second example the helicopter has to go through moving openings in two walls: this example can be particularly difficult for randomized planners, since the environment is characterized by the presence of "narrow passages". The basic algorithm, as reported in

[10], was not able to handle this case very effectively. However, the extensions presented in this paper resulted in a significant reduction in the required computation time: in our tests the average time required to compute a feasible solution was only 2.1 seconds (standard deviation 1.3 seconds).

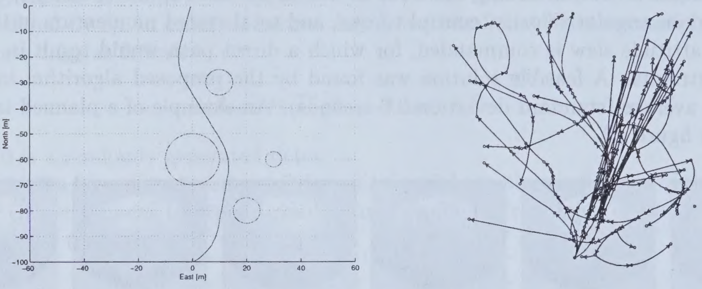


Figure 3: Helicopter flying among fixed spheres: planned trajectory (left) and trace of the trajectory tree (right)

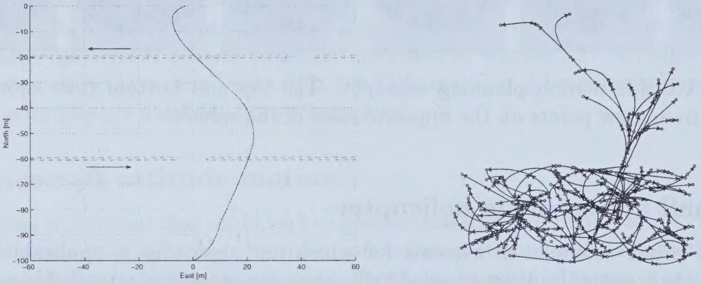


Figure 4: Helicopter flying through moving doors: planned trajectory (left) and trace of the trajectory tree (right)

5 Conclusions

In this paper we have presented a randomized motion planning algorithm, based on the use of Lyapunov functions as local planners in a probabilistic roadmap framework. The main advantage of the algorithm is the capability to address in an efficient and natural fashion the dynamics of the system, while at the same time providing a consistent decoupling between the motion planning and the low-level control tasks. We have introduced extensions to a previously published version of the algorithm, which resulted in significant improvements in its performance. The resulting algorithm has been shown to be flexible enough to handle system evolving on several kinds of state space, including manifolds and hybrid state spaces.

6 Acknowledgements

The authors would like to thank Dr. Richard Kornfeld, of the Jet Propulsion Laboratory, for helpful discussion on the spacecraft attitude motion planning problem. This research was supported by the C. S. Draper Laboratory through the IR&D grant DL-H-505334, by the AFOSR, under grant F49620-99-1-0320, and by the ONR, under Young Investigator Award N-00014-99-1-0668.

References

[1] J.-C. Latombe. Motion planning: a journey of robots, molecules, digital actors, and other artifacts. *International Journal of Robotics Research*, 18(11), November 1999.

[2] J.C. Latombe. *Robot Motion Planning.* Kluwer Academic Publishers, Boston, MA, 1991.

[3] Z. Li and J.F. Canny, editors. *Nonholonomic Motion Planning.* Kluwer Academic Publishers, Boston, MA, 1993.

[4] J.-P. Laumond, editor. *Robot Motion Planning and Control*, volume 229 of *Lectures Notes in Control and Information Sciences.* Springer, 1998. Available at `http://www.laas.fr/ jpl/book.html` at the time of writing.

[5] L.E. Kavraki, P. Svestka, J.C. Latombe, and M.H. Overmars. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Transactions on Robotics and Automation*, 12(4):566–580, 1996.

[6] J. Barraquand and J.C. Latombe. Nonholonomic multibody mobile robots: controllability and motion planning in the presence of obstacles. *Algorithmica*, 10(2-4):121–155, 1993.

[7] J. E. Marsden and T.S. Ratiu. *Introduction to Mechanics and Symmetry.* Springer Verlag, New York, NY, 1999. Second edition.

[8] E. Frazzoli, M.A. Dahleh, and E. Feron. A hybrid control architecture for aggressive maneuvering of autonomous helicopters. In *IEEE Conf. on Decision and Control*, December 1999.

[9] E. Frazzoli, M.A. Dahleh, and E. Feron. Robust hybrid control for autonomous vehicle motion planning. Technical Report LIDS-P-2468, Massachusetts Institute of Technology, 1999. Revised version submitted to IEEE Trans. on Automatic Control.

[10] E. Frazzoli, M.A. Dahleh, and E. Feron. Real-time motion planning for agile autonomous vehicles. In *AIAA Conf. on Guidance, Navigation and Control*, Denver, CO, August 2000.

[11] D. Hsu, J.C. Kindel, J.-C. Latombe, and S. Rock. Randomized kinodynamic motion planning with moving obstacles. In *Proc. Workshop on Algorithmic Foundations of Robotics (WAFR'00)*, Hanover, NH, March 2000.

[12] S.M. LaValle and J.J Kuffner. Randomized kinodynamic planning. In *Proceedings of the 1999 IEEE International Conference on Robotics and Automation*, 1999.

[13] F. Bullo. *Nonlinear Control of Mechanical Systems: A Riemannian Geometry Approach.* PhD thesis, California Institute of Technology, 1998.

[14] T.J. Koo and S. Sastry. Output tracking control design of a helicopter model based on approximate linearization. In *IEEE Conference on Decision and Control*, 1998.

[15] E. Frazzoli, M.A. Dahleh, and E. Feron. Trajectory tracking control design for autonomous helicopters using a backstepping algorithm. In *American Control Conference*, Chicago, IL, 2000.

Using Randomization to Find and Optimize Feasible Trajectories for Nonlinear Systems

Peng Cheng Zuojun Shen Steven M. LaValle

Department of Computer Science
Iowa State University
Ames, IA 50011 USA
{chp, zjshen, lavalle}@iastate.edu

Abstract

We present our current progress on the design and experimentation with trajectory planning and optimization algorithms for nonlinear systems that have significant state-space constraints. An overview of our planning method based on Rapidly-exploring Random Trees (RRTs) is given. We show our current planning results for two challenging sets of nonlinear systems: 1) the determination of automobile trajectories through obstacle courses; 2) the design of re-entry trajectories for a reusable launch vehicle model based on the NASA X33 prototype. We also briefly describe some early results on using randomization to optimize trajectories in the presence of state space constraints.

1 Introduction

The design of open loop trajectories for nonlinear systems that have state space constraints is a challenging task. If algorithms can be designed that are able to efficiently find and optimize such trajectories for broad classes of problems, many application areas would be greatly impacted, including robotics, aeronautics, and automotive design. The successful development of such algorithms will most likely depend on a culmination of ideas from both the algorithmic motion planning and nonlinear systems communities.

The primary interest in the motion planning community has been to compute collision free paths in the presence of complicated constraints on the configuration space of one or more controllable bodies (while ignoring differential motion constraints). It is widely known that the class of problems is NP-hard [33], which has caused the focus of research in this area to move from exact, complete algorithms to randomized (or Monte-Carlo) algorithms that can solve many challenging high-dimensional problems efficiently at the expense of completeness. Within the past decade, randomized versions of earlier ideas were developed. The classic notion of a *roadmap* [6, 20, 31], is a network of collision-free paths that captures the configuration-space topology, and is generated by preprocessing the configuration space independently of any initial-goal query. This evolved into a Monte-Carlo variation termed a probabilistic roadmap (PRM) [16], which is formed by selecting numerous configurations at random, and generating a network of paths by attempting to connect nearby points. In contrast to roadmaps, classical *incremental search* ideas are based heavily on a particular initial-goal query, and include methods such as dynamic programming, A^* search, or bidirectional search. These evolved into randomized approaches such as the randomized potential field approach [1], Ariadne's clew algorithm [28], the planner in [15][1], and Rapidly-exploring Random Trees (RRTs) [18].

Once differential constraints are introduced, a challenging problem emerges that involves both nonlinear control and traditional path planning issues. This problem is often referred to as *nonholonomic planning* [2, 21, 12, 22, 30] or kinodynamic planning [5, 7, 8, 10, 9, 12]. Many of these methods, such as those in [2, 9], follow closely the incremental search paradigm. It is generally more challenging to design a roadmap-based algorithm due to the increased difficulty of connecting numerous pairs of states in the presence of differential constraints (often referred to as the steering problem [22]). The first randomized approach to kinodynamic planning appeared in [24], and was based on Rapidly-exploring

---

[1] We note that the method introduced here is termed a PRM by the authors. Since the method is based on incremental search, we include it here to help categorize the methods based on their conceptual similarities.

Random Trees [23]. In that paper, RRTs were applied to trajectory design problems for hovercrafts and rigid spacecrafts that move in a cluttered 2D or 3D environment. Theoretical performance bounds are presented in [25]. In [13], RRTs were applied to the design of collision-free trajectories for a helicopter in a cluttered 3D environment. In [34], RRTs were applied to the design of trajectories for underactuated vehicles. Recently, the incremental search method in [17] was extended to the case of kinodynamic planning in time-varying environments. Sections 3-5 present the RRT-based trajectory design method and its application to vehicle dynamics and a reusable launch vehicle.

Given the emerging interest in the use of randomized techniques for trajectory design, a growing issue is that the quality of the resulting trajectories may be poor due to the large number of random fluctuations in inputs. This has led us recently to consider the problem of trajectory refinement as a post-processing step to our planning methods. In Section 6 our current attempts to use randomized ideas in this domain are briefly presented.

2 Problem Description

The class of problems considered in this paper can be formulated in terms of six components:

1. **State Space:** A topological space, X
2. **Boundary Values:** $x_{init} \in X$ and $X_{goal} \subset X$
3. **Collision Detector:** A function, $D : X \to \{true, false\}$, that determines whether global constraints are satisfied from state x. This could alternatively be a real-valued function.
4. **Inputs:** A set, U, which specifies the complete set of controls or actions that can affect the state.
5. **Incremental Simulator:** Given the current state, $x(t)$, and inputs applied over a time interval, $\{u(t')|t \leq t' \leq t + \Delta t\}$, compute $x(t + \Delta t)$.
6. **Metric:** A real-valued function, $\rho : X \times X \to [0, \infty)$, which specifies the distance between pairs of points in X (however, ρ is not necessarily symmetric).

Trajectory planning will generally be viewed as a search in a state space, X, for a control u that brings the system from an initial state, x_{init} to a goal region $X_{goal} \subset X$ or goal state $x_{goal} \in X$. It is assumed that a complicated set of global constraints is imposed on X, and any solution path must keep the state within this set. A collision detector reports whether a given state, x, satisfies the global constraints. We generally use the notation X_{free} to refer to the set of all states that satisfy the global constraints. Local, differential constraints are imposed through the definition of a set of inputs (or controls) and an incremental simulator. Taken together, these two components specify possible changes in state. The incremental simulator can be defined by numerical integration of a *state transition equation* of the form $\dot{x} = f(x, u)$, or can simply be achieved by a simulation software package. Finally, a metric is defined to indicate the closeness of pairs of points in the state space. This metric will be used in Section 3, when the RRT is introduced.

3 Trajectory Planning Method

The Rapidly-exploring Random Tree (RRT) was introduced in [23] as a planning algorithm to quickly search high-dimensional spaces that have both algebraic constraints (arising from obstacles) and differential constraints (arising from kinematics and dynamics). The key idea is to bias the exploration toward unexplored portions of the space by sampling points in the state space, and incrementally "pulling" the search tree toward them.

The basic RRT construction algorithm is given in Figure 1. A simple iteration is performed in which each step attempts to extend the RRT by adding a new vertex that is biased by a randomly-selected state, $x \in X$. The EXTEND function, illustrated in Figure 2, selects the nearest vertex already in the RRT to x. The "nearest" vertex is chosen according to the metric, ρ. The function NEW_STATE makes a motion toward x by applying an input $u \in U$ for some time increment Δt. This input can be chosen at random, or selected by trying all possible inputs and choosing the one that yields a new state as close as possible to the sample, x (if U is infinite, then an approximation or analytical technique can be used).

```
BUILD_RRT($x_{init}$)
    $\mathcal{T}$.init($x_{init}$);
    for $k = 1$ to $K$ do
        $x_{rand} \leftarrow$ RANDOM_STATE();
        EXTEND($\mathcal{T}, x_{rand}$);
    Return $\mathcal{T}$
```

```
EXTEND($\mathcal{T}, x$)
    $x_{near} \leftarrow$ NEAREST_NEIGHBOR($x, \mathcal{T}$);
    if NEW_STATE($x, x_{near}, x_{new}, u_{new}$) then
        $\mathcal{T}$.add_vertex($x_{new}$);
        $\mathcal{T}$.add_edge($x_{near}, x_{new}, u_{new}$);
```

Figure 1: The basic RRT construction algorithm.

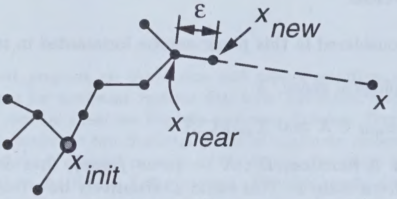


Figure 2: The EXTEND operation.

NEW_STATE implicitly uses the collision detection function to determine whether the new state (and all intermediate states) satisfy the global constraints. For many problems, this can be performed quickly ("almost constant time") using incremental distance computation algorithms [14, 26, 29] by storing the relevant invariants with each of the RRT vertices. If NEW_STATE is successful, the new state and input are represented in x_{new} and u_{new}, respectively. The left column of Figure 3 shows an RRT for a 2D holonomic (no differential constraints) planning problem, constructed in a 2D square space. The incremental construction method biases the RRT to rapidly explore in the beginning, and then converge to a uniform coverage of the space [25].

One issue is the size of the step that is used for RRT construction. This could be chosen dynamically during execution on the basis of a distance computation function that is used for collision detection. If the bodies are far from colliding, then larger steps can be taken. Aside from following this idea to obtain an incremental step, how far should the new state, x_{new} appear from x_{near}? Should we try to connect x_{near} to x_{rand}? Instead of attempting to extend an RRT by an incremental step, EXTEND can be iterated until the random state is reached or until the collision detector reports a collision. The best choice often depends on the type of problem.

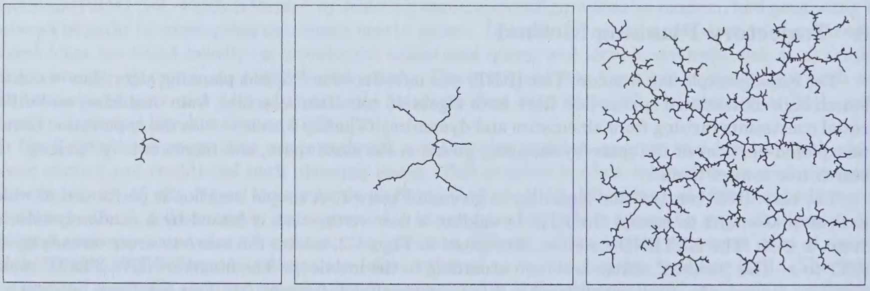


Figure 3: The RRT rapidly explores in the beginning, before converging to the sampling distribution.

The simple RRT can be used in a variety of ways to form path planning methods [25]. The recommended choice depends on several factors, such as whether differential constraints exist, the type of collision detection algorithm, or the efficiency of nearest neighbor computations. By changing the sampling distribution from uniform a biased RRT can be constructed. For example, an RRT can be started at x_{init}, and samples from X_{goal} can be occasionally chosen in the place of random samples. Alternatively, a gradual bias of samples around the goal region can be performed. Inspired by classical bidirectional search techniques [32], it seems reasonable to expect that improved performance can be obtained by growing two RRTs, one from x_{init} and the other from x_{goal}; a solution is found if the two RRTs meet. This divides the computation time between two processes: 1) exploring the state space; 2) trying to grow the trees into each other. Two trees, $\mathcal{T}_a$ and $\mathcal{T}_b$ are maintained at all times until they become connected and a solution is found. In each iteration, one tree is extended, and an attempt is made to connect the nearest vertex of the other tree to the new vertex. Then, the roles are reversed by swapping the two trees.

4 Trajectory Design for Automobiles

Simulators are used extensively to evaluate vehicle performance in the automotive industry. Most of these simulations involve close interaction between a human driver and a simulation system. The experiments presented in this section can be considered as a component that might be used in virtual prototyping of automobile designs. As opposed to requiring human interaction, the RRT-based planner serves as a kind of "virtual stunt driver" that attempts to achieve specified conditions. The nine-dimensional model used here (and others we have used in our experiments) are minor variations of the models considered in [3].

Variables and constants The nine-dimensional state vector is $(x, y, r, \psi, \phi, q, \nu, s, \beta)$. The 3D coordinate frame is designed with the x coordinate increasing from left to right, the y coordinate increasing from top to bottom, and the z coordinate inward (to form a right-handed coordinate system). Let β be the steering angle. Let a and b be the distance from the front and rear axles to the car center, respectively. Let ψ be the yaw angle of the car. Let s be the forward speed of the car, and let ν be the sideways speed (arising from slipping). Let r be the angular velocity. Let ϕ be the roll (which describes the sideways tilting of the car). Let q be the roll angle rate. Let α_f and α_r be the slipping angle of the front and rear wheels, respectively. These are expressed as

$$\alpha_f = \frac{\nu + ar}{s} - \beta$$

and

$$\alpha_r = \frac{\nu - br}{s}.$$

Let $C_{\alpha f}$ and $C_{\alpha r}$ be the cornering stiffness between the forces along the y axis, F_{yf} and F_{yr}, respectively, on the front and rear wheels. Under some conditions, it is possible for the car to slip sideways. If $N_f\mu/2 > C_{\alpha f}\tan(|\alpha_f|)$, the calculated friction force is less than the maximum possible friction, then $F_{yf} = -C_{\alpha f}\alpha_f$; otherwise, $F_{yf} = \mu N_f Sgn(\alpha_f)(1 - x_f/2)$, in which Sgn denotes the sign function, μ is a constant, and $x_f = N_f\mu/2C_{\alpha f}\tan(|\alpha_f|)$. Similarly, if $N_r\mu/2 > C_{\alpha r}\tan(|\alpha_r|)$, then $F_{yr} = -C_{\alpha r}\alpha_r$; otherwise, $F_{yr} = \mu N_r Sgn(\alpha_r)(1 - x_r/2)$, in which $x_r = N_f\mu/2C_{\alpha r}\tan(|\alpha_r|)$. Let M be the car mass, and let I be the yaw moment of inertia. Let H_2 be the distance from the joint connecting the chassis with the car frame (the chassis and frame are flexibly attached to model a simple suspension system). The constants K, c, and other details are described an in [3].

Equations of motion In the equations, let $h = (-(K - MgH_2)\phi - cq - (F_{yf} + F_{yr})H_2)/I$. The following represent the nine equations of motion: $\dot{x} = s\cos\psi - \nu\sin\psi$, $\dot{y} = s\sin\psi + \nu\cos\psi$, $\dot{r} = (F_{yf}a - F_{yr}b)/I$, $\dot{\psi} = r$, $\dot{\phi} = q$, $\dot{q} = h$, $\dot{\nu} = (F_{yf} + F_{yr})/M - sr - H_2 h$, $\dot{s} = u_1$, $\dot{\beta} = u_2$.

The inputs are u_1, which is linear acceleration, and u_2, which is the change in the steering angle.

Experiments Figure 4 shows three separate experiments that were performed for the car model using a simple RRT with a goal bias (the goal was chosen with probability 1/20, instead of a random sample). Figure 4.a involves a lane-changing problem, which is referred to in the automotive industry as the Consumer Union Short Course. The car is moving at 60 m.p.h. from left to right, and planner generates

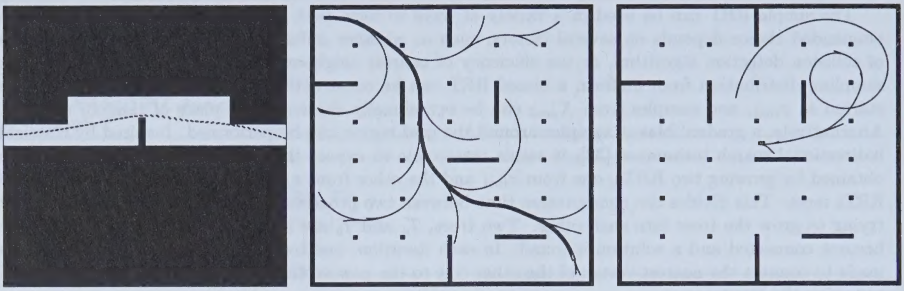

(a) (b) (c)

Figure 4: a) Double lane change; b) fast driving through an obstacle coursel c) forcing the tires to lift.

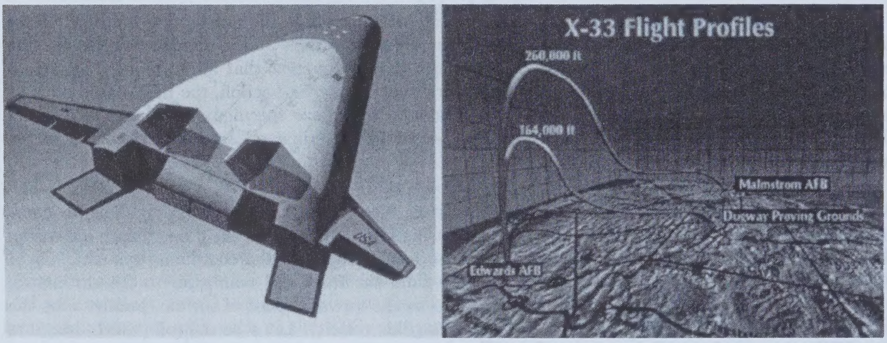


Figure 5: The NASA X-33 prototype reusable launch vehicle.

a steering input that avoids collision. Figure 4.b shows an example in which a car must travel at 108 k.p.h. through an obstacle course. To further complicate the problem, the load on all four tires was computed using the models in [3], and a state constraint was defined that prevents any tire from lifting off of the ground. In Figure 4.c, the car starts at 18 k.p.h., and is requested to reach a state that causes a tire to lift from the ground. The car executes a spiraling path to generate sufficient speed, and then turns while going 190 k.p.h. to achieve the goal. Such a test could be used, for example, in a vehicle safety study.

5 Trajectory Design for a Reusable Launch Vehicle

One of the current goals in space transportation is to design a new generation of launch vehicles to dramatically lower the costs of putting payloads in space. Instead of the complex systems in use today, the emphasis is being placed on a simple, fully reusable vehicle. Since July 1996, NASA has commissioned Lockheed Martin Skunk Works to design, build and test the X-33 experimental vehicle, which is depicted in Figure 5. The purpose of the vehicle is to serve as a prototype for demonstrating the feasibility of the new concept. The craft is capable of reaching speeds beyond Mach 13. A formidable challenge is to design a trajectory that will guide the X-33 safely to earth. Skilled engineers sometimes spend weeks designing good trajectories for this system. In this section, we describe our early, but promising results in the design of trajectories for the X-33.

Variables and constants Our model is a minor variation of the one presented in [27]. The equations of motion will be expressed in terms of dimensionless variables. In the basic model there are six state variables, $r, \theta, \phi, V, \gamma$, and ψ. Let r be the radial distance from the center of the earth to the flying spacecraft, normalized by the radius of the earth, $R_0 = 6378km$. The longitude and latitude are θ and ϕ, respectively. Derivatives of these variables will be taken with respect to dimensionless time τ, in which $\tau = \frac{t}{\sqrt{R_0 g_0}}$ and $g_0 = 9.81m/sec^2$. Let V be the earth-relative velocity, normalized by $\sqrt{R_0 g_0}$. Let γ be the flight path angle measured from horizon surface downward. Let ψ be the velocity azimuth angle, measured from the north in a clockwise direction.

The equations of motions will include the variables D, L, and Ω, which are all functions of state. The quantities D and L are aerodynamic accelerations in g's, which are expressed as

$$L = \frac{1}{2}\rho V^2 S C_L \tag{1}$$

$$D = \frac{1}{2}\rho V^2 S C_D, \tag{2}$$

in which S is a constant based on surface area, and $C_L = h_1(M, \alpha)$ and $C_D = h_2(M, \alpha)$ are the lift and drag coefficients, respectively. These are each functions of Mach number, M, and the angle of attack, α. The quantity ρ is the air density, which is a function of r. The quantity α could be pre-scheduled to be a function of M, or α could alternatively be used as a control variable. For simplicity, we assume it is a function of M. The functions, h_1 and h_2 are defined using table lookup and interpolation.

Let Ω be the rotation rate of the earth normalized by $\sqrt{R_0 g_0}$. Let σ be the bank angle, which can be considered as the principle control input of the re-entry flight (the actual control surface inputs on the craft are computed in terms of bank angle).

Equations of motion The dimensionless equations of three dimensional motion over a spherical rotating earth are:

$$\dot{r} = V \sin\gamma \tag{3}$$

$$\dot{\theta} = \frac{V \cos\gamma \sin\psi}{r \cos\phi} \tag{4}$$

$$\dot{\phi} = \frac{V \cos\gamma \cos\psi}{r} \tag{5}$$

$$\dot{V} = -D - \left(\frac{\sin\gamma}{r^2}\right) + \Omega^2 r \cos\phi(\sin\gamma\cos\phi - \cos\gamma\sin\phi\cos\psi) \tag{6}$$

$$\dot{\gamma} = \frac{1}{V}\left\{L\cos\sigma + \left(V^2 - \frac{1}{r}\right)\left(\frac{\cos\gamma}{r}\right) + 2\Omega V\cos\phi\sin\psi + \Omega^2 r\cos\phi(\cos\gamma\cos\phi - \sin\gamma\cos\psi\sin\phi)\right\} \tag{7}$$

$$\dot{\psi} = \frac{1}{V}\left[\frac{L\sin\sigma}{\cos\gamma} + \frac{V^2}{r}\cos\gamma\sin\psi\tan\phi - 2\Omega V(\tan\gamma\cos\psi\cos\phi - \sin\phi) + \frac{\Omega^2}{\cos\gamma}\sin\psi\sin\phi\cos\phi\right]. \tag{8}$$

Boundary conditions Initial state is given by orbit flight conditions. The goal state is termed the TAEM (Terminal Area Energy Management) point, which is:

$$r(\tau_f) = r_f, \qquad \theta(\tau_f) = \theta_f, \qquad \phi(\tau_f) = \phi_f \tag{9}$$

$$V(\tau_f) = V_f, \qquad \gamma_{min} \le \gamma(\tau_f) \le \gamma_{max}, \qquad \psi(\tau_f) = \psi_f \tag{10}$$

State constaints The following constraints define X_{free}. Although the X-33 is not in "collision" in the physical sense, failure to satisfy these state-space constraints will result in a collision that must be reported by the collision detector, as described in Section 2.

Normal load constraint:

$$|L \cos \alpha + D \sin \alpha| \leq n_{z_{max}} \tag{11}$$

Dynamic pressure constraint:

$$q \leq q_{max} \tag{12}$$

Heat rate constraint:

$$\dot{Q}_s \leq \dot{Q}_{max} \tag{13}$$

Equilibrium glide constraint:

$$\left[\frac{1}{r} - V^2\right]\left(\frac{1}{r}\right) - L \leq 0 \tag{14}$$

Both $\dot{Q}_s$ and q are complicated functions of state. A slice of the resulting constraints is shown in Figure 6.a.

Varying models We have performed experiments with several variations of the model. Each is progressively more complicated than the previous one, due to a difference in the input.

A The state is $(r, \theta, \phi, V, \gamma, \psi)$, and the input is the bank angle, σ.

B The state is $(r, \theta, \phi, V, \gamma, \psi, \sigma)$, and the input is bank angle velocity, $\dot{\sigma}$.

C The state is $(r, \theta, \phi, V, \gamma, \psi, \sigma, \nu)$, and the input is bank angle acceleration, $\ddot{\sigma}$. The new state variable is defined as $\nu = \dot{\sigma}$.

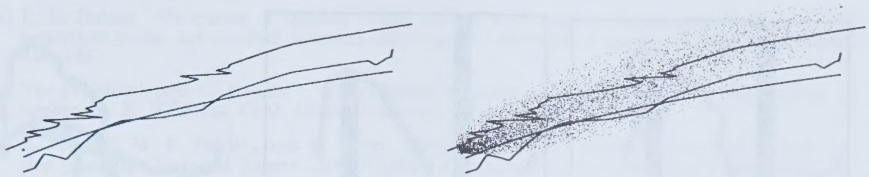
Model A allows instantaneous setting of the bank angle, which is unrealistic in practice. Model B allows the bank angle velocity to be changed. Model C is the most realistic (and most difficult), because only the acceleration of the bank angle is provided by the input. The limits on the bank angle derivatives are $|\dot{\sigma}| \leq 10^o/sec$ and $|\ddot{\sigma}| \leq 5^o/sec^2$.

Experiments In our experiments to date, we have experienced considerable success design trajectories using RRT-based planners on Models A and B. In both cases we are able to reach the TAEM. Currently, we are experimenting with Model C. Figure 6 shows a computed trajectory for Model B. The initial state is $r = 1.0086$, $\theta = 4.257$, $\phi = 0.6400$, $V = 0.3630$, $\gamma = 0.01098$, $\psi = 0.6138$, $\sigma = 0.0$. The goal state is $r = 1.0045$, $\theta = 4.305$, $\phi = 0.6967$, $V = 0.1156$, $\gamma = -0.1062$, $\psi = 0.7522$, $\sigma = -0.2250$. A weighted Euclidean metric was used in the RRT algorithm. Figure 6.a shows the projection of the state space constraints, plotted as speed vs. range. The state must remain below the jagged upper curve, and above both of the two lower curves (although the initial state may not initially be below the upper curve). The initial and goal states are indicated with black dots on the right and left of the figure, respectively. Figure 6.b shows a projection of the random samples that were used in the RRT. Instead of using uniform random samples, the samples were concentrated in an elliptical region based on the flight corridor, with some bias towards the goal state. Figures 6.c-f show the paths generated by the RRT. In each case, the initial and goal states are shown as dots. The resulting trajectory comes sufficiently close to the goal state in each of the perspectives shown.

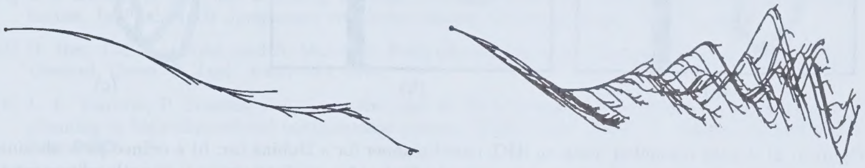
6 Path Smoothing and Optimization

Although randomization is helpful in the design of feasible trajectories, it often leads to jagged paths or useless fluctuations in the inputs. This motivations our investigation into the problem of refining trajectories that are computed by our planning algorithms. Existing optimization methods, such as first-order gradient descent [4], are usually not designed to handle complicated, implicit state-space constraints that usually emerge in motion planning, and also make differentiability assumptions on the model (which would might prohibit their use in the models used in Sections 4 and 5.

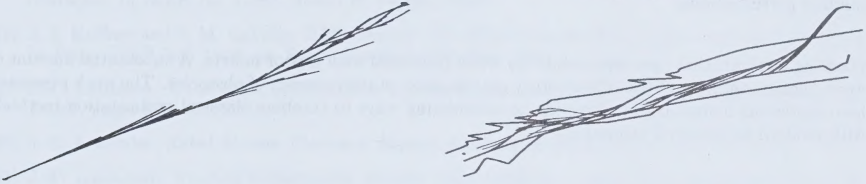
We are currently exploring the use of randomization to refine trajectories in the presence of state space constraints. Some experiments are shown here for the Dubins car [11]. In this model, there are three state variables (position and orientation). The car moves forward-only at constant speed; the



(a) Flight corridor: Altitude vs. speed (b) Biased sampling along flight corridor



Altitude vs. range Heading angle vs. range



Latitude vs. longitude Altitude vs. speed

Figure 6: Trajectory design results for the X-33 reusable launch vehicle.

only input is the steering angle. The maximum steering angle induces a bound on the curvature of the trajectory.

Two methods are currently being evaluated. For the first method, consider the classical first-order gradient method presented in [4]. For a given control history, $u(t)$ for $t \in [0, t_f)$, this iterative method yields a perturbation, $u(t) + \delta u(t)$, that locally (in the trajectory space) reduces a performance criterion. Eventually, convergence to a locally-optimal solution is obtained; however, the basic method becomes quickly trapped if obstacles exist. One way to avoid this problem is to iteratively select a segment of the path at random, perform the classical optimization on it, and then insert the new segment into the original path. Figure 7.a shows a path computed by an RRT-based planner, and Figure 7.b shows the resulting path after performing a few hundred iterations of the classical gradient descent method on randomly-selected path segments. We note that Dubins' shortest path curves [11, 19] could be used instead of the method in [4]; however, this approach cannot be generalized to many systems. An alternative method is to iteratively compute $\delta u(t)$, and use any perturbation $u(t) + \delta u(t)$ that satisfy collision constraints and reduce the performance criterion. This method can be particularly useful if the equations of motion are not differentiable or not even explicitly known. Figure 7.c shows an initial path, two intermediate paths, and a final optimized path that was obtained using random perturbations. We have yet to consider obstacles in this method, and to compare both methods. Nevertheless, it appears so far that randomization may by useful for path smoothing and optimization.

7 Discussion

The experiments presented here show good promise for the use of randomized planning and optimization methods for challenging problems that involve nonlinear systems and complicated state-space constraints. However, many challenges remain. One of the greatest difficulties in motion planning of trajectories is designing a metric that yields good performance. For some systems is may be possible to utilize cost-to-go or Lyapunov functions; however, it also appears valuable to develop randomized

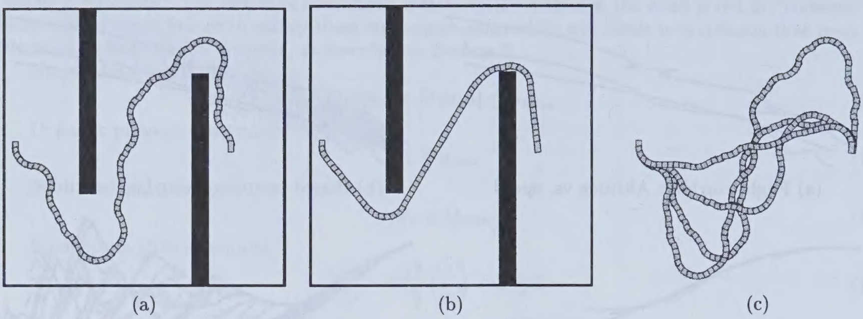


Figure 7: a) A path computed using an RRT-based planner for a Dubins car; b) a refined path obtained by iteratively optimizing random intervals using the first-order gradient method; c) path refinement by random perturbations.

planning methods that have less sensitivity when presented with a poor metric. A substantial amount of work remains on the problem of trajectory optimization in the presence of obstacles. The work presented here represents a step in that direction be considering ways to combine classical optimization methods with modern randomized algorithms.

Acknowledgments

This work was funded in part by NSF CAREER Award IRI-9875304 (LaValle). We thank Jim Bernard for many helpful discussions on vehicle dynamics, and his suggestions of models. We thank Ping Lu for his suggestions regarding the modeling of the X-33. We thank Francesco Bullo for helpful discussions regarding trajectory optimization.

References

[1] J. Barraquand, B. Langlois, and J. C. Latombe. Numerical potential field techniques for robot path planning. *IEEE Trans. Syst., Man, Cybern.*, 22(2):224–241, 1992.

[2] J. Barraquand and J.-C. Latombe. Nonholonomic multibody mobile robots: Controllability and motion planning in the presence of obstacles. *Algorithmica*, 10:121–155, 1993.

[3] J. Bernard, J. Shannan, and M. Vanderploeg. Vehicle rollover on smooth surfaces. In *Proc. SAE Passenger Car Meeting and Exposition*, Dearborn, Michigan, 1989.

[4] A. E. Bryson and Y.-C. Ho. *Applied Optimal Control.* Hemisphere Publishing Corp., New York, NY, 1975.

[5] J. Canny, A. Rege, and J. Reif. An exact algorithm for kinodynamic planning in the plane. *Discrete and Computational Geometry*, 6:461–484, 1991.

[6] J. F. Canny. *The Complexity of Robot Motion Planning.* MIT Press, Cambridge, MA, 1988.

[7] M. Cherif. Kinodynamic motion planning for all-terrain wheeled vehicles. In *IEEE Int. Conf. Robot. & Autom.*, 1999.

[8] C. Connolly, R. Grupen, and K. Souccar. A Hamiltonian framework for kinodynamic planning. In *Proc. of the IEEE International Conf. on Robotics and Automation (ICRA'95)*, Nagoya, Japan, 1995.

[9] B. R. Donald and P. G. Xavier. Provably good approximation algorithms for optimal kinodynamic planning for Cartesian robots and open-chain manipulators. *Algorithmica*, 14(6):480–530, 1995.

[10] B. R. Donald, P. G. Xavier, J. Canny, and J. Reif. Kinodynamic planning. *Journal of the ACM*, 40:1048–66, November 1993.

[11] L. E. Dubins. On curves of minimal length with a constraint on average curvature, and with prescribed initial and terminal positions and tangents. *American Journal of Mathematics*, 79:497–516, 1957.

[12] Th. Fraichard and C. Laugier. Kinodynamic planning in a structured and time-varying 2d workspace. In *IEEE Int. Conf. Robot. & Autom.*, pages 2: 1500–1505, 1992.

[13] E. Frazzoli, M. A. Dahleh, and E. Feron. Robust hybrid control for autonomous vehicles motion planning. Technical Report LIDS-P-2468, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, 1999.

[14] L. J. Guibas, D. Hsu, and L. Zhang. H-Walk: Hierarchical distance computation for moving convex bodies. In *Proc. ACM Symposium on Computational Geometry*, pages 265–273, 1999.

[15] D. Hsu, J.-C. Latombe, and R. Motwani. Path planning in expansive configuration spaces. *Int. J. Comput. Geom. & Appl.*, 4:495–512, 1999.

[16] L. E. Kavraki, P. Svestka, J.-C. Latombe, and M. H. Overmars. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Trans. Robot. & Autom.*, 12(4):566–580, June 1996.

[17] R. Kindel, D. Hsu, J.-C. Latombe, and S. Rock. Kinodynamic motion planning amidst moving obstacles. In *IEEE Int. Conf. Robot. & Autom.*, 2000.

[18] J. J. Kuffner and S. M. LaValle. RRT-connect: An efficient approach to single-query path planning. In *Proc. IEEE Int'l Conf. on Robotics and Automation*, 2000.

[19] J.-C. Latombe. A fast path planner for a car-like indoor mobile robot. In *Proc. Am. Assoc. Artif. Intell.*, pages 659–665, 1991.

[20] J.-C. Latombe. *Robot Motion Planning.* Kluwer Academic Publishers, Boston, MA, 1991.

[21] J.-P. Laumond. Finding collision-free smooth trajectories for a non-holonomic mobile robot. In *Proc. Int. Joint Conf. on Artif. Intell.*, pages 1 20–1123, 1987.

[22] J. P. Laumond, S. Sekhavat, and F. Lamiraux. Guidelines in nonholonomic motion planning for mobile robots. In J.-P. Laumond, editor, *Robot Motion Plannning and Control*, pages 1–53. Springer-Verlag, Berlin, 1998.

[23] S. M. LaValle. Rapidly-exploring random trees: A new tool for path planning. TR 98-11, Computer Science Dept., Iowa State University. <http://janowiec.cs.iastate.edu/papers/rrt.ps>, Oct. 1998.

[24] S. M. LaValle and J. J. Kuffner. Randomized kinodynamic planning. In *Proc. IEEE Int'l Conf. on Robotics and Automation*, pages 473–479, 1999.

[25] S. M. LaValle and J. J. Kuffner. Rapidly-exploring random trees: Progress and prospects. In *Workship on the Algorithmic Foundations of Robotics*, 2000.

[26] M. C. Lin and J. F. Canny. Efficient algorithms for incremental distance computation. In *IEEE Int. Conf. Robot. & Autom.*, 1991.

[27] P. Lu and J. M. Hanson. Entry guidance for the X-33 vehicle. *J. Spacecraft and Rockets*, 35(3):342–349, 1998.

[28] E. Mazer, G. Talbi, J. M. Ahuactzin, and P. Bessière. The Ariadne's clew algorithm. In *Proc. Int. Conf. of Society of Adaptive Behavior*, Honolulu, 1992.

[29] B. Mirtich. V-Clip: Fast and robust polyhedral collision detection. Technical Report TR97-05, Mitsubishi Electronics Research Laboratory, 1997.

[30] R. M. Murray and S. Sastry. Nonholonomic motion planning: Steering using sinusoids. *Trans. Automatic Control*, 38(5):700–716, 1993.

[31] C. O'Dunlaing and C. K. Yap. A retraction method for planning the motion of a disc. *Journal of Algorithms*, 6:104–111, 1982.

[32] I. Pohl. Bi-directional and heuristic search in path problems. Technical report, Stanford Linear Accelerator Center, 1969.

[33] J. H. Reif. Complexity of the mover's problem and generalizations. In *Proc. of IEEE Symp. on Foundat. of Comp. Sci.*, pages 421–427, 1979.

[34] G. J. Toussaint, T. Başar, and F. Bullo. Motion planning for nonlinear underactuated vehicles using hinfinity techniques. Coordinated Science Lab, University of Illinois, September 2000.

Motion Planning for a Class of Hybrid Locomotion Systems

Miloš Žefran
Electrical Engineering and Computer Science
University of Illinois at Chicago
Chicago, IL 60607
mzefran@eecs.uic.edu

Abstract

A mechanical system subject to constraints typically undergoes changes in the dynamics as the set of active constraints changes, and can be thus modeled as a hybrid system. The focus of this paper is on devices that switch between constraints in order to locomote. An important component of a motion plan for a hybrid system is a sequence of dynamic regimes that it needs to traverse to reach the goal. Finding this sequence can lead to a combinatorial complexity of motion planning algorithms. We show that controllability analysis can be exploited to reduce this complexity. However, since switches between dynamic regimes result in discontinuous changes in the vector fields describing the system, classical controllability results are not applicable. We characterize the controllability of locomotion devices that are able to switch between dynamic regimes at an arbitrary point in the configuration space. We use the analysis to compute sub-optimal motion plans for a simple locomotion device consisting of a planar linkage sliding on a frictionless surface that can clamp a subset of its links to the surface.

1 Introduction

In this paper we study motion planning for a class of systems that can locomote by switching between constraints. Locomotion devices belong to the class of mechanical control systems (Lagrangian systems) and have a special structure that can be fruitfully exploited for control. However, several locomotion modalities, including legged locomotion involve impacts and switches in the dynamic behavior and are therefore inherently non-smooth. Non-smooth locomotion systems belong to the class of hybrid systems. While a number of approaches to modeling and control of hybrid systems exist [1, 6], these works are in general not applicable to locomotion. Some exceptions are [5] which provides a quasi-static controllability analysis of locomotion devices, and [15] where a method for stabilizing systems with changing dynamics is described.

The state space of a non-smooth locomotion system can be partitioned into regions D_j so that in the region D_j the system is described with a set of differential equations $\dot{x} = F_j(x, u)$. The regions D_j are called discrete states; within each discrete state, the differential equations describe the evolution of the continuous state. Let $\gamma(x, u, t)$ be a trajectory that connects the state x_0 with x_1 (Figure 1). This trajectory is characterized

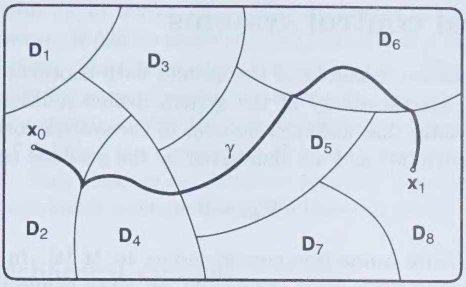


Figure 1: A trajectory of a system with discrete and continuous state.

by a sequence of time instants $T_0 = t_0 < t_1 < \ldots < t_{N+1} = T_1$ and a sequence of indices $j_0, \ldots, j_N$ such that on the interval $[t_i, t_{i+1}]$, the trajectory belongs to the set D_{j_i} and at the time t_{i+1} it switches from the region D_{j_i} to $D_{j_{i+1}}$. A motion plan for a hybrid locomotion system thus consists of three components:

(i) A sequence, $\{D_{j_i}\}_{i=0}^{N}$, of discrete states.

(ii) A sequence, $\{t_i\}_{i=1}^{N}$, of switching times.

(iii) Continuous trajectories on each interval $[t_i, t_{i+1}]$.

We have shown in our previous work [16] that if the sequence of discrete states is known, motion plans can be efficiently computed using variational approaches. However, finding a sequence of discrete states involves descending the graph describing the discrete evolution of the system and can easily lead to a combinatorial explosion. In [14] we showed that for a certain class of hybrid locomotion systems, the sequence of discrete states necessary to move in a particular direction can be inferred from the controllability analysis of the system. The aim of the present paper is therefore to put together these two sets of results and demonstrate how the analytical tools from [14] can be used in numerical motion planning algorithms such as [16].

The main assumption that allows us to perform the controllability analysis is that the system can switch between constraints at any point in the configuration space. In other words, the distribution that describes the set of admissible directions for a particular constraint is defined at every point. An example is a planar linkage sliding on a frictionless surface that can arbitrarily clamp a subset of its links to the surface. Note that the above assumption does not hold for legged robots; there, the switch to a constrained regime only takes place when a leg hits the ground (the constrained distribution is only defined on a submanifold).

The underlying mathematical framework of this paper is Riemannian geometry as it applies to mechanical control systems [11]. Modeling and controllability results for smooth mechanical systems are discussed in [10]. Of special importance in the controllability analysis are series describing the evolution of the system [12, 13]. Such series are also a basis of several motion planning algorithms [2, 7, 8].

We base the paper on the modeling framework for constrained mechanical systems outlined in [9]. We define a class of hybrid mechanical control systems and show that under the appropriate assumptions the controllability results from [9] can be extended to this class of systems. Finally, we show how the controllability analysis can be used in motion planning.

2 Mechanical control systems

Let Q be the configuration manifold of the system with coordinates $q = (q_1, \dots, q_n)$. At every $q \in Q$, the kinetic energy of the system defines a Riemannian metric, M_q. Hamilton's principle states that unforced motions of the system correspond to geodesics with respect to the metric M[1] and are thus given by the geodesic equation:

$$\nabla_{\dot{q}}\dot{q} = 0 \tag{1}$$

where ∇ is the *Levi-Civita connection* corresponding to M [4]. In coordinates (we use the summation convention throughout the paper), an affine connection is given by:

$$\nabla_X Y = \left(\frac{\partial Y^i}{\partial q^j} X^j + \Gamma^i_{jk} X^j Y^k \right) \frac{\partial}{\partial q^i} \tag{2}$$

where the coefficients Γ^i_{jk} are known as Christoffel symbols.

When an external force (a one-form) F acts on the system, the dynamic equations take the form:

$$\nabla_{\dot{q}}\dot{q} = Y \tag{3}$$

where $Y(q) = M_q^{-1}F(q)$ is now a vector field. Written in coordinates, these *forced Euler-Lagrange* equations take the familiar form:

$$\ddot{q} + M_q^{-1}C(q,\dot{q}) = M_q^{-1}F \tag{4}$$

where M_q is the inertia matrix and $C(q,\dot{q})$ are the Coriolis and centrifugal forces. Formally a *mechanical control system* can be defined as a tuple $(Q, M, \mathcal{F}, U)$, where Q is an n-dimensional configuration manifold, M is a Riemannian metric on Q (the kinetic energy), $\mathcal{F} = \text{span}\{F^1, \dots, F^m\}$ is the input co-distribution (input forces), and $U \subset \mathbb{R}^m$ is the set of inputs. Later in the paper we will also need the notion of input distribution, $\mathcal{Y} = \text{span}\{M^{-1}F^i\}$.

2.1 Constraints

¿From the control point of view, constraints on a mechanical system limit the set of directions in which the system can move. Therefore, an intrinsic description of a constraint is a distribution on Q, describing at each point the set of feasible velocities. Such description applies both to holonomic and nonholonomic constraints.

A mechanical control system together with a constrained distribution will be called a *constrained mechanical control system*, $\Sigma = (Q, M, \mathcal{F}, \mathcal{D}, U)$. It was shown in [9] that the dynamic equations for such a system can be also written using an affine connection:

$$\widetilde{\nabla}_{\dot{q}}\dot{q} = u^k P_{\mathcal{D}}(Y^k), \tag{5}$$

where $P_{\mathcal{D}} : TQ \to TQ$ is the orthogonal projection to $\mathcal{D}$ (with respect to the metric M) and $\widetilde{\nabla}$ is given by

$$\widetilde{\nabla}_X Y = \nabla_X Y + A^{-1}\left(\nabla_X A(I-P)\right)(Y), \qquad \forall X, Y$$

---

[1] The controllability analysis is simpler if the potential forces are present so we will neglect them in this work.

where A is an invertible $(1,1)$ tensor. For different choices of A we will get different affine connections, however it can be easily checked that if Y is a vector field which takes values in D then all these connections yield the same value. This implies that each of these connections yields the same dynamic equations on the constraint distribution $\mathcal{D}$. The tensor A is usually chosen so that the computations are simplified.

The importance of defining the connection $\widetilde{\nabla}$ is that equation (5) is formally identical to equation (3). Such description therefore provides a unified treatment of both constrained and unconstrained mechanical control systems, see [9].

2.2 Hybrid mechanical control systems

A *hybrid mechanical control system* consists of a mechanical control system $(Q, M, \mathcal{F}, U)$ together with a given set of constraint distributions $\mathcal{D}_i$, where i belongs to an index set I. Each constraint $\mathcal{D}_i$ yields a constrained mechanical control system $\Sigma_i = (Q, M, \mathcal{F}, U, \mathcal{D}_i)$, with associated affine connection ∇_i and input distribution $\mathcal{Y}_i = \operatorname{span}\{P_{\mathcal{D}_i} M^{-1} F\}$. Formally, the hybrid mechanical control system is therefore a tuple $(I, Q, M, \mathcal{F}, U, \{\mathcal{D}_i\}_{i \in I})$. Slightly more general definition can be found in [3].

The evolution of a hybrid mechanical control system can be described as follows. The system starts in a state $((q, \dot{q}), i) \in TQ \times I$ and it evolves according to the dynamics given by ∇_i and the chosen set of controls. At any point, we can choose to switch to any other discrete state. Whenever the system switches between two discrete states i and j (constraint distributions $\mathcal{D}_i$ and $\mathcal{D}_j$) it undergoes impact. The velocity after the impact is the orthogonal projection onto $\mathcal{D}_j$ of the velocity before the impact:

$$\dot{q}(t^+) = P_{\mathcal{D}_j} \dot{q}(t^-) \tag{6}$$

3 Equilibrium controllability

In the case of mechanical control systems two controllability questions can be asked: (a) what points in the *phase space* (i.e., configuration and velocity) can be reached; and (b) what *configurations* can be reached. The first question can be addressed using standard nonlinear controllability methods. We will show that in the case of hybrid mechanical control systems, the second question is particularly interesting. For smooth systems, the question was posed and answered in [10] and we briefly review this work here.

Consider a system described by (5). Let q_0 be a point in Q and let W be a neighborhood of q_0. The *reachable set of* q_0 *within* W is

$$\mathcal{R}_Q^W(q_0, \leq T) = \bigcup_{\tau \leq T} \{x \in Q \mid \exists \text{ a solution } q(t) \text{ to (5) s.t. } \dot{q}(0) = 0,\ q(t) \in W \text{ for } t \in [0, \tau],\ q(\tau) = x\}.$$

Note that the definition of $\mathcal{R}_Q^W(q_0, \leq T)$ only involves the configurations, not the velocities. The system (5) is *locally configuration controllable* at q_0 if there exists a time T such that $\mathcal{R}_Q^W(q_0, \leq T)$ contains a neighborhood of q_0 for any neighborhood W of q_0, and *equilibrium controllable* on $W \subset Q$, if for any two equilibrium points $q_1, q_2 \in W$, there exists an input $\{u^k(t), t \in [0, T]\}$ and a solution $\{q(t), t \in [0, T]\}$ such that $q(0) = q_1$, $q(T) = q_2$, $q(t) \in W$ for all $t \in [0, T]$, and $\dot{q}(0) = 0$, $\dot{q}(T) = 0$.

To characterize the configuration controllability, we introduce the following operations. As in [10], we define the *symmetric product* of two vector fields as:

$$\langle X : Y \rangle = \nabla_X Y + \nabla_Y X.$$

Let $\mathcal{X} = \{X_1, \dots, X_m\}$ be a family of vector fields. We define $\mathrm{Lie}(\mathcal{X})$ to be the closure of $\mathcal{X}$ under the Lie bracket operation (the involutive closure), and we let $\mathrm{Sym}(\mathcal{X})$ be the closure of $\mathcal{X}$ under the symmetric product operation. Within the set $\mathrm{Sym}(\mathcal{X})$, we define the *order* of a symmetric product to be the number of vector fields X_j present in it. We say that a symmetric product is *bad* if it contains an even number of each X_i. Otherwise the product is said to be *good*. The controllability tests are then:

Theorem 3.1 ([10]). *The system is configuration controllable at* q_0 *if:*

(i) the rank of $\mathrm{Lie}(\mathrm{Sym}(\mathcal{Y}))$ *is full;*

(ii) at q_0*, every bad symmetric product is a linear combination of lower order good symmetric products.*

If these conditions are verified at every $q \in W$*, then the system is equilibrium controllable on* W*.*

3.1 Hybrid mechanical control systems

In [3] we studied configuration and equilibrium controllability for hybrid mechanical control systems. Since the analysis above requires computation of the brackets at zero velocity, we had to restrict our analysis to the case when the switches between different regimes occur at zero velocity. We showed:

Proposition 3.2 ([3]). *A hybrid mechanical control system is equilibrium controllable on an open set* W *if the following two conditions hold:*

(i) in each discrete state i, every bad symmetric product is a linear combination of lower order good symmetric products

(ii) the rank of $\mathrm{Lie}(\sum_{i \in I} \mathrm{Sym}_i(\mathcal{Y}_i))(q)$ *is full for all* $q \in W$*.*

Restricting a hybrid mechanical control system to switches at zero velocity is overly restrictive. It was shown in [14] that it is possible to generalize the above result to switches at nonzero velocity. The key observation is that if in some regime i the system can generate a nonzero velocity at q_0 (by making a small loop in the configuration space), then we can use the nonzero velocity to exploit directions generated by the brackets that would otherwise vanish. These additional directions can not be generated by starting at zero velocity and must be evaluated for those values of v that can be generated by making small loops from q_0. These directions are characterized by $\mathrm{Sym}(\mathcal{Y})$, see [10].

For simplicity we assume that the system only has two regimes (the system can switch between free motion and motion in directions $\mathcal{D}$), $\Sigma = (\{1,2\}, Q, M, \mathcal{F}, U, \{TQ, \mathcal{D}\})$.

Proposition 3.3 ([14]). *The system* Σ *is equilibrium controllable on some neighborhood* W *of* q_0 *if the following conditions hold:*

(i) $\mathcal{D} \subseteq \mathrm{span}\,\{\mathrm{Sym}(\{Y_i^1\}) \cup \mathrm{Sym}(\{Y_i^2\})\}$*, where* $Y_i^1 = Y_i$ *and* $Y_i^2 = P_{\mathcal{D}} Y_i^1$*.*

(ii) For $k = 1, 2$*, the bad symmetric products* $\langle Y_i^k : Y_i^k \rangle$ *are spanned by the vector fields* Y_i^k*.*

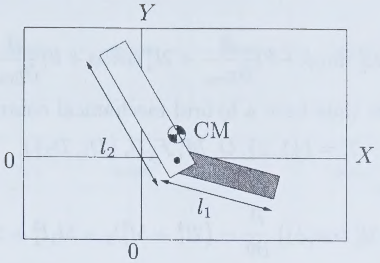


Figure 2: Sliding and clamping mechanism with two links. One of the links can be clamped to the floor so that it is completely immobilized.

(iii) For every $q \in W$:

$$TQ = \operatorname{span}\left\{\mathcal{D} \cup \{Y_i^k, [Y_i^k, Y_j^l]\} \cup \{\nabla_v Y_i^k, \nabla^2 Y_i^k(v,v) + R(v, Y_i^k)v\}_{v \in \mathrm{Sym}(\{Y_i^l\})}\right\}$$

where k and l must be different regimes.

Condition (ii) corresponds to condition (i) in Proposition 3.2, while condition (iii) is similar to condition (ii) there with added terms that can be generated by switches at nonzero velocity. These terms are evaluated on the set of velocities that can be generated in regime 2. Conditions (ii) and (iii) imply only local configuration controllability, which means that we would not necessarily be able to stop after we moved to the desired configuration. However, if condition (i) holds, we can bring the velocity to zero at the final position by performing loops there. The system will be therefore equilibrium controllable.

4 Controllability analysis and planning for a sliding and clamping mechanism

In this section we apply our analysis to a two-link mechanism sliding without friction on a plane. The mechanism consists of two homogeneous bars of unit density and lengths (l_1, l_2), connected by an actuated rotational joint (Figure 2). In the figure, CM denotes the center of mass of the two body system. The coordinates of the center of mass of the link j are (x_j, y_j, θ_j), while $(x_{\mathrm{CM}}, y_{\mathrm{CM}})$ are the coordinates of CM. One of the links can be instantaneously clamped to the ground (anywhere on the plane) so that it gets completely immobilized. When the link is clamped, the number of degrees of freedom of the system decreases from 4 to 1. The clamping constraint is described by the function $\varphi_1(q) = (x_1, y_1, \theta_1)$.

The configuration manifold of the two body system is $Q = SE(2) \times S^1$. We will describe the configuration with the coordinates $q = (x_{\mathrm{CM}}, y_{\mathrm{CM}}, \theta, \phi)$, where $\theta = \theta_1$ and $\phi = \theta_2 - \theta_1$.

When the first link is clamped, the system is confined to the submanifold $R_1(q_0) = \{q \in Q \mid \varphi_1(q) = (x_0, y_0, \theta_0)\}$. This holonomic constraint induces the constraint distribution:

$$\mathcal{D}_2(q) = \operatorname{span}\left\{-2l_2^2\sin(\phi+\theta)\frac{\partial}{\partial x_{\mathrm{CM}}} + 2l_2^2\cos(\phi+\theta)\frac{\partial}{\partial y_{\mathrm{CM}}} + 4(l_1+l_2)\frac{\partial}{\partial \phi}\right\}$$

If we set $\mathcal{D}_1(q) = T_qQ$, we thus have a hybrid mechanical control system:

$$\Sigma = (\{1,2\}, Q, M, \mathcal{F}, U, \{\mathcal{D}_1, \mathcal{D}_2\}).$$

The input vector field is:

$$Y_1^1 = l_2^2\left(l_2(5l_1+2l_2) + 3l_1^2\cos(\phi)\right)\frac{\partial}{\partial\theta} - \left(2l_1^4 + 5l_1^3l_2 + 5l_1l_2^3 + 2l_2^4 + 6l_1^2l_2^2\cos(\phi)\right)\frac{\partial}{\partial\phi}$$

The input vector field on the constrained regime is obtained by projecting Y_1^1 onto the constraint distribution. Since the constrained distribution is one-dimensional, the constrained input spans exactly the same direction as the constrained distribution:

$$Y_1^2 = -2l_2^2\sin(\phi+\theta)\frac{\partial}{\partial x_{\mathrm{CM}}} + 2l_2^2\cos(\phi+\theta)\frac{\partial}{\partial y_{\mathrm{CM}}} + 4(l_1+l_2)\frac{\partial}{\partial\phi}$$

We now have all the necessary tools to check for equilibrium controllability. Notice that in the regime 2 the system is fully controllable, $\mathrm{Sym}(\{Y_1^2\}) = \mathcal{D}$, so the condition (i) is trivially satisfied. We can also show that $\langle Y_1^1 : Y_1^1\rangle = \zeta Y_1^1$ for some scalar function ζ, so that condition (ii) is satisfied. Finally, we can check that

$$\operatorname{rank}\left\{Y_1^1, Y_1^2, [Y_1^1, Y_1^2], \nabla^2 Y_1^1(Y_1^2, Y_1^2) + R(Y_1^2, Y_1^1)Y_1^2\right\}(q) = 4$$

in a neighborhood of the point $(x_{\mathrm{CM}}, y_{\mathrm{CM}}, \theta, \phi) = (0,0,0,0)$. Therefore, the hybrid mechanical control system Σ is equilibrium controllable.

It turns out that we could replace the last vector with the bracket $[Y_1^1, [Y_1^2, Y_1^1]]$ and still span the whole space. This means that the system is equilibrium controllable even if we only allow switches at zero velocity, as was shown in [3]. However, this bracket is of higher order than the bracket generated by the switches at nonzero velocity. Therefore, even if switching at nonzero velocity does not contribute any new controllable directions, it allows us to move in certain directions more efficiently.

The controllability analysis also indicates what switching sequence will be necessary to move (locally) in a particular direction. Clearly, the directions Y_1^1 and Y_1^2 do not require any switches. The direction $[Y_1^1, Y_1^2]$ can be generated by a sequence *UCUC*, where *U* stands for the unclamped regime and *C* for the clamped regime. And finally, the direction $\nabla^2 Y_1^1(Y_1^2, Y_1^2) + R(Y_1^2, Y_1^1)Y_1^2$ will require the sequence *CUC*. In the clamped regime we first generate a non-zero velocity in the direction Y_1^2. After switching to the unclamped regime the system will start moving in the direction $\nabla^2 Y_1^1(Y_1^2, Y_1^2) + R(Y_1^2, Y_1^1)Y_1^2$, but in order to stop the motion we need to switch to the clamped regime again at the end of the maneuver.

At this point a motion planning method such as that described in [16] can be employed to construct a motion plan. Since the controllability analysis only gives us local results, we will in general need to concatenate a series of maneuvers in order to reach the desired goal configuration. Figure 3 shows for example a maneuver that can move the mechanism for a unit distance in the y direction. The mechanism is initially clamped. It then swings the unclamped link to build the velocity, after which it releases the constraint and starts drifting. At the end, a maneuver is performed that stops the mechanism.

The motion plan in the figure was computed using the sequential quadratic programming method provided in the NAG numerical library. For the simulation, we fixed the duration of motion to 3s, with the switches occuring at 1s and 2s. The computed trajectory minimizes the L^2 norm of the input.

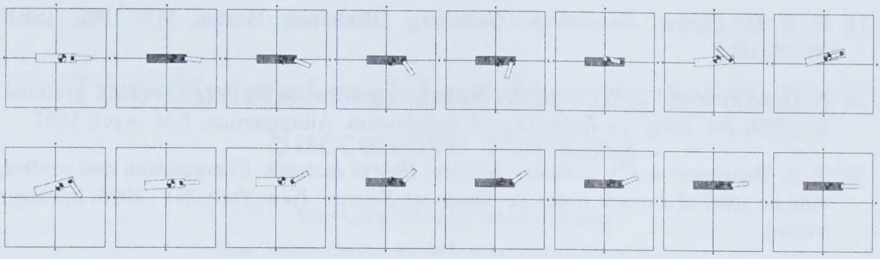


Figure 3: Motion generated by switch at nonzero velocity. When the link is clamped to the ground, it is colored in gray.

5 Conclusions

We investigated a class of mechanisms that can locomote by switching between constraints. The non-smooth nature of such systems due to changes in the dynamic equations and impacts prevents the application of conventional tools for motion planning and control. We showed that under certain conditions it is possible to perform the controllability analysis for such non-smooth locomotion devices by exploiting their special Lagrangian structure. The controllability analysis leads to motion primitives that can be efficiently computed and subsequently used to construct motion plans. We also characterized motion directions in the case when the system switches between constraints at nonzero velocity. We showed that switches at nonzero velocity can provide new controllable directions and can generate motions in certain directions more efficiently than gaits provided by the zero velocity analysis.

Acknowledgments

The authors would like to thank Jim Radford for helpful discussions. This research was supported by the National Science Foundation under grants CMS-9502224 and CSE-9704702.

References

[1] P. Antsaklis, W. Kohn, A. Nerode, and S. S. Sastry, editors. *Hybrid systems IV*, volume 1273 of *Lecture notes in Computer Science*, New York, NY, 1997. Springer Verlag.

[2] F. Bullo, N. Ehrich Leonard, and A. D. Lewis. Controllability and motion algorithms for underactuated Lagrangian systems on Lie groups. *IEEE Transactions on Automatic Control*, 45(8):1437–1454, 2000.

[3] F. Bullo and M. Žefran. On modeling and locomotion of hybrid mechanical systems with impacts. In *IEEE Conf. on Decision and Control*, pages 2633–2638, Tampa, FL, December 1998.

[4] M. P. Do Carmo. *Riemannian Geometry.* Birkhäuser, Boston, MA, 1992. ISBN 0817634908.

[5] B. Goodwine and J. W. Burdick. Trajectory generation for legged robotic systems. In *IEEE Int. Conf. on Robotics and Automation*, Albuquerque, NM, April 1997.

[6] T. A. Henzinger and S. S. Sastry, editors. *Hybrid systems: Computation and control*, volume 1386 of *Lecture notes in Computer Science*, New York, NY, 1998. Springer Verlag.

[7] G. Lafferriere and H. J. Sussmann. A differential geometric approach to motion planning. In Z. Li and J. F. Canny, editors, *Nonholonomic Motion Planning*, pages 235–270. Kluwer Academic Publishers, 1993. ISBN 0792392752.

[8] N. Ehrich Leonard and P. S. Krishnaprasad. Motion control of drift-free, left-invariant systems on Lie groups. *IEEE Transactions on Automatic Control*, 40 (9):1539–1554, 1995.

[9] A. D. Lewis. Simple mechanical control systems with constraints. *IEEE Transactions on Automatic Control*, 45(8):1420–1436, 2000.

[10] A. D. Lewis and R. M. Murray. Configuration controllability of simple mechanical control systems. *SIAM Journal on Control and Optimization*, 35(3):766–790, 1997.

[11] J. E. Marsden and T. S. Ratiu. *Introduction to Mechanics and Symmetry.* Springer Verlag, New York, NY, second edition, 1999. ISBN 038798643X.

[12] H. J. Sussmann. A product expansion of the Chen series. In C. I. Byrnes and A. Lindquist, editors, *Theory and Applications of Nonlinear Control Systems*, pages 323–335. Elsevier, Oxford, UK, 1986. ISBN 0444700552.

[13] H. J. Sussmann. A general theorem on local controllability. *SIAM Journal on Control and Optimization*, 25(1):158–194, 1987.

[14] M. Žefran, F. Bullo, and J. Radford. An investigation into non-smooth locomotion. In *IEEE Int. Conf. on Robotics and Automation*, pages 2038–2043, Detroit, MI, May 1999.

[15] M. Žefran and J. W. Burdick. Design of switching controllers for systems with changing dynamics. In *IEEE Conf. on Decision and Control*, Tampa, FL, December 1998.

[16] M. Žefran, J. Desai, and V. Kumar. Continuous motion plans for robotic systems with changing dynamic behavior. In *Robotic motion and manipulation*, pages 113–128. A K Peters, Wellesley, MA, 1997.

Two-dimensional Interleaving by Using Successive Packing

Yun-Qing Shi and Ximin Zhang

New Jersey Center for Wireless Telecommunication
Dept. of Electrical and Computer Engineering
New Jersey Institute of Technology
Newark, NJ 07102
shi@njit.edu

Abstract

In this paper, we present a new two-dimensional (2-D) interleaving method, called successive packing (SP), to combat 2-D burst errors. It is shown that in a 2-D array of $2^n \times 2^n$, interleaved with the SP method, any error burst of $2^k \times 2^k$ $(1 \leq k < n)$, or $2^k \times 2^{k+1}$ or $2^{k+1} \times 2^k$ $(0 \leq k < n)$ can be corrected by using some simple random error correction codes. The method is optimal for the above mentioned set of bursts in the sense that the interleaving degree achieves the lower bound. A brief performance comparison with some existing techniques is made and some future research is discussed.

1 Introduction

One-dimensional (1-D) interleaving technique has been included in some communication and error correction coding texts, say, in [1]. In many applications such as magnetic and optical data storage (including holographic storage), charge-coupled devices (CCDs), 2-D barcodes and information hiding in digital images, 2-D bursts of errors do take place. Therefore, the protection of 2-D information has arisen as an important task that is of both theoretical and practical importance.

In the 2-D barcode application, the United Parcel Service (UPS) uses the 1-D interleaving technique to mix up code symbols and then follows some pattern, say, the Boustrophedonic pattern to write the 1-D interleaved code symbol sequence into the 2-D barcodes. Conceivably, the 1-D interleaving technique cannot work effectively for correcting 2-D burst errors [2]. Recently some 2-D burst error correction codes have been developed [3], [4], [5], [6], [2]. Note that developing multi-dimensional (M-D) interleaving techniques in order to combat burst errors with some simple random-error-correcting codes has become the most common approach to the task [7], [3], [2]. Among all of these results, that presented in [3] is the most recent and comprehensive one. The method is implemented according to the size of a burst of errors, t (the number of the code symbols contained in the burst). When t increases, a different algorithm has to be implemented to produce a different interleaved array to combat the larger burst. When the actual size of a burst is less than t, the interleaved array is not optimal anymore. On the contrary, in many applications, the burst size may not be known exactly in advance, while the size of 2-D array is known. Under the circumstances, we propose a novel 2-D interleaving technique. It can correct bursts with a set of different sizes and hence can be implemented only once. With respect to these bursts, it is optimal in the sense that the interleaving degree achieves the lower bound, meaning that the smallest possible number of codewords is required in the interleaving scheme.

2 Successive Packing for 2-D Interleaving

The strategy of 2-D interleaving to combat 2-D bursts of errors is similar to that of the 1-D interleaving. That is, the elements of a given 2-D array are mixed up in the interleaved array such that the elements of a burst are separated away enough from one another. If in the de-interleaved 2-D array there is at most a single error in one codeword, then the 2-D burst can be corrected by using a one-random-error-correcting code. In this paper, we assume one-random-error-correcting codes for the sake of brevity. It is understood that the results can be extended to r-random-error-correcting codes (with integer $r > 1$).

To apply the SP, the subscripts of all the elements in a 2-D array, that needs to be interleaved, should be manipulated as follows. Considering a 2-D array, that consists of a certain number of codewords. In general, they are 2-D codewords. 1-D codewords (row-wise or column-wise) can be considered as a special case of 2-D codewords. For instance, if the length of codewords is 4, then the codeword may be a 1×4 sub-array, or a 4×1 sub-array, or a 2-D sub-array of 2×2, or a sub-array with a different shape, say, a bent-line, as shown in Figure 1. We need to number all code symbols within a codeword so as to make their subscripts form a 1-D subsequence. All 1-D subsequences of this type need to be cascaded to form a composite 1-D sequence. In this way, we convert all code symbols in the 2-D array into a 1-D sequence of code symbols before applying the SP.

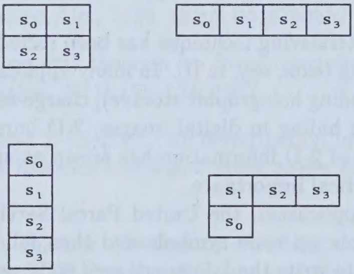


Figure 1: The different types of elements block: square, row, column, bent line.

Procedure 2.1. *The 2-D interleaving using the successive packing proceeds as follows. Consider a 2-D array of* $2^n \times 2^n$.

When $n = 0$, *the interleaved array of* 1×1 *is the original array itself. That is,*

$$S_1 = [s_0] \tag{1}$$

where s_0 *represents the element in the array, and* S_1 *the interleaved array with its subscript representing the total number of elements in it. In general, for a given* n, *the interleaved array is denoted by* $S_{2^{2n}}$.

The procedure is carried out successively. That is, for a given interleaved array S_i, the interleaved array of S_{4i} can be generated according to

$$S_{4i} = \begin{bmatrix} 4 \times S_i + 0 & 4 \times S_i + 2 \\ 4 \times S_i + 3 & 4 \times S_i + 1 \end{bmatrix} \tag{2}$$

where the notation of $4 \times S_i + k$ $(k = 0, 1, 2, 3)$ represents a 2-D array that is generated from S_i. This means that $4 \times S_i + k$ has the same dimensionality as S_i. Furthermore, each element in $4 \times S_i + k$ is indexed in such a way that its subscript equals to the four times of that of the corresponding (having the same position) elements in S_i plus k.

From the elements' indexes point of view, Equation 2 indicates that S_{4i} is derived from S_i by *packing* S_i four times. This is where the term successive packing comes from. According to the above rule, we have

$$S_4 = \begin{bmatrix} 4 \times S_1 + 0 & 4 \times S_1 + 2 \\ 4 \times S_1 + 3 & 4 \times S_1 + 1 \end{bmatrix} = \begin{bmatrix} s_0 & s_2 \\ s_3 & s_1 \end{bmatrix} \tag{3}$$

Similarly, we have S_{16} as follows.

$$S_{16} = \begin{bmatrix} 4 \times S_4 + 0 & 4 \times S_4 + 2 \\ 4 \times S_4 + 3 & 4 \times S_4 + 1 \end{bmatrix} = \begin{bmatrix} s_0 & s_8 & s_2 & s_{10} \\ s_{12} & s_4 & s_{14} & s_6 \\ s_3 & s_{11} & s_1 & s_9 \\ s_{15} & s_7 & s_{13} & s_5 \end{bmatrix} \tag{4}$$

Prior to presenting the main result contained in Theorem 2.1, we introduce the following prerequisite.

Definition 2.1. *Consider a 2-D array of $2^n \times 2^n$, the subscripts of all of whose elements form a 1-D sequence in some fashion. Partition the array into 2^m blocks with the integer m satisfying $1 \leq m \leq 2n - 1$. Each block thus generated contains $K = 2^{2n-m}$ elements. That is, each block is of size $K = 2^{2n-m}$. The first block contains elements having the subscripts from 0 to $2^{2n-m} - 1$. In general, for any element having its subscript k satisfying $2^{2n-m}(d-1) \leq k < 2^{2n-m}d$, $d \leq 2^m$, we say that this element belongs to the dth block. All the elements belonging to the same block are said to be K-equivalent to one another, and form a K-equivalent element class.*

Definition 2.2. *Consider two bursts, B_1 and B_2, in an interleaved 2-D array. If these two bursts have the same size and shape, and each element in a burst (say, B_1) is either an element of another burst (say, B_2) or a K-equivalent element of an element of another burst (say, B_2), and vice versa, then we call these two bursts, B_1 and B_2, the K-equivalent bursts.*

Dealing with burst error correction, we may consider each block defined in Definition 2.1 as a 2-D codeword. A burst in the interleaved array is said to be corrected if each element in the burst has been spread in the de-interleaved array into a distinct codeword. From this point of view, it is easy to see that given two equivalent bursts, if one can be interleaved and corrected, then the other can also be interleaved and corrected.

Lemma 2.1. *Let A be an array of $2^n \times 2^n$ interleaved by using the successive packing procedure. Then, any burst of $2^k \times 2^k$ in A, with the integer k satisfying $1 \leq k < n$, are the K_1-equivalent to one another, where $K_1 = 2^{2n-2k}$; and any burst of $2^k \times 2^{k+1}$ or $2^{k+1} \times 2^k$ in A, with the integer k satisfying $0 \leq k < n$, are the K_2-equivalent to one another, where $K_2 = 2^{2n-2k-1}$.*

Proof. We prove the Lemma for the square bursts first followed by a proof for the rectangular bursts. In either cases, the mathematical induction method is employed. Since in either case the proof is trivial as $n = 0, 1$, and is very simple as $n = 2$, we start the proof as $n = 3$ in either case.

I1. When $n = 3$, A is of 8×8, as shown in Figure 2 (a). $k < n$ implies that $k = 0, 1, 2$. Since the proof for $k = 0$ is trivial, we only prove the Lemma for $k = 1, 2$.

I2. Consider $k = 2$. There are 16 elements (code symbols) in each of this type of square bursts. To correct this type of the bursts with a one-random-error-correcting code, the size of the codewords is 4. Look at the burst of 4×4, located at the top-left quadrant of A. In particular, look at the very northwest corner element, s_0. According to Definition 2.1, its 4-equivalent elements are: s_1, s_2, s_3, which are marked in Figure 2 (a). Similarly, each element in the burst has its own 4-equivalent elements, which follow the same distribution pattern as the 4-equivalent elements of s_0. Therefore, if we consider that the burst of 4×4, located at the top-left quadrant of A, is shifted toward the right by one column, we observe that the burst and its shifted version have three columns in common, while the first column in the burst (consisting of the elements $s_0, s_{48}, s_{12}, s_{60}$) has been replaced by the first column in the top-right quadrant (consisting of $s_2, s_{50}, s_{14}, s_{62}$). It is easy to see that each element in the first column of the burst is 4-equivalent to an element in the first column in the top-right quadrant. Hence, these two bursts are 4-equivalent according to Definition 2.2.

I3. This reasoning can be used for three more times (each time the burst of 4×4 is shifted towards the right by one column) and we can then conclude that the 4×4 burst located in the top-left quadrant is equivalent to each of its right-shifted versions.

I4. Applying the reasoning, contained in I2 and I3, accordingly, we arrive at that the burst of 4×4, located in the top-left quadrant in A, is 4-equivalent to each of its down-shifted versions. After applying the same reasoning, contained in I2 and I3, to all the down-shifted burst of 4×4 described in the previous sentence in considering its equivalence with its right-shifted versions, we conclude that any burst of 4×4 in A are 4-equivalent to each other. Thus we conclude the proof of the Lemma for any square bursts when $k = 2$.

I5. Consider $k = 1$. There are 4 elements in each of this type of square bursts. To correct this type of burst errors with a one-random-error-correcting code, the size of the codewords is 16. Look at the very northwest 2×2 burst in A. In particular, the very northwest corner element s_0. Figure 2 (a) shows all its 16-equivalent elements. Applying the reasoning used in the proof for $k = 2$ accordingly leads to the completion of the proof for $k = 1$.

I6. Having finished the proof for $n = 3$, we assume now that the Lemma holds for any square bursts in the A when $n = N$. (Hence, A may be denoted by $S_{2^{2N}}$.) To complete the proof for the Lemma to hold for any square bursts, according to the mathematical induction, we only need to show that the Lemma holds for any square bursts in A when $n = N + 1$. (Note that at this time A may be denoted by $S_{2^{2N+2}}$.)

I7. To complete the proof for square bursts as $n = N + 1$, we prove the Lemma first for $k < N$, then for $k = N$.

I8. Consider $k < N$. When $n = N$, each element in the very northwest square burst of $2^k \times 2^k$ has $2^{2N-2k} - 1$ equivalent elements in $S_{2^{2N}}$ due to the assumption made in I6. When $S_{2^{2N+2}}$ is successively packed by using $S_{2^{2N}}$ according to Equation 2, the total number of the equivalent elements of each element in the very northwest square burst of $2^k \times 2^k$, located in the top-left quadrant of $S_{2^{2N+2}}$, has increased by four times. Meanwhile, the difference in the subscripts of any two equivalent elements in the top-left quadrant of $s_{2^{2N+2}}$ has increased by four time compared with that in the $s_{2^{2N}}$. The increased equivalent elements are distributed evenly among the other three quadrants of $s_{2^{2N+2}}$ due to the successive packing (refer to Equation 2). It is therefore easy to see that the equivalent elements of any specific elements in the very northwest square burst of $2^k \times 2^k$ occupy the same positions in each of the four quadrants of $s_{2^{2N+2}}$. Furthermore, these positions are the same as that in the $s_{2^{2N}}$.

Obviously, this observation is true for any elements in the very northwest burst of $2^k \times 2^k$. Therefore, the same reasoning used to prove the case of 8×8 can be applied accordingly to complete the proof for $k < N$.

I9. When $k = N$, the total number of the equivalent elements of each elements in the square burst located in the top-left quadrant is 4. Hence the situation is similar to the square bursts of 4×4 when $n = 3$. We can use the reasoning, contained in I2, to prove that the Lemma holds for $k = N$. Up to this point we have proved the Lemma for any square bursts in A.

We now prove the Lemma for any rectangular bursts defined in Lemma 2.1.

II1. Consider $n = 3$. All the six possible types of rectangular bursts described in the Lemma in this situation are: 1×2, 2×4, 4×8, 2×1, 4×2, and 8×4.

II2. Consider the rectangular bursts of 4×8 located at the top-half of the 8×8 array. Under the circumstances, the total number of the equivalent elements of any element in the rectangular burst is 2. The equivalent elements of s_0 and s_2 are s_1 and s_3 respectively. If two equivalent elements are linked with a line segments, it is found that these two line segments cross each other. It is noted, in general, that the line segments linking any element in the left-half of the burst of 4×8, located at the top-half of the 8×8 array, with its equivalent element will have the same orientation as that of the line segment linking s_0 and its equivalent element s_1, while the line segment linking any element in the right-half of the burst with its equivalent element has the same orientation with that linking s_2 and its equivalent element s_3. Obviously, the burst of 4×8 at the top-half in A is equivalent to its one-row-down-shifted version. Repeat this procedure and use the transitive property of the equivalence relation, we can conclude that the burst is equivalent to any of its down-shifted version. This implies that all the bursts of 4×8 are equivalent to each other.

II3. Now prove the Lemma for any burst of 2×4. At this time, the dimensionality of the equivalence is 8. Take a look at such a burst, which is located at the very northwest corner of the interleaved array of 8×8. All the 8-equivalent elements of the element s_0, together with s_0, are shown in Figure 2 (b) with the solid line segments linking them. It is seen that these eight elements form two zigzag patterns in the array. All the 8-equivalent elements of the element s_8, including the element s_8, are shown in Figure 2 (b) with the dashed line segments linking them. Note that the zigzag pattern associated with s_8 has just "opposite" orientation to that with s_0. It is easy to verify that all elements in the left-half of the burst of 2×4 have an equivalent element distribution pattern similar to that of s_0, while all elements in the right-half of the burst have an equivalent element distribution pattern similar to that of s_8.

Once this feature is identified, the reasoning used in the previous proof for the square bursts when $n = 3$ can be used accordingly to finish the proof for any burst of 2×4. For any burst of 1×2, the dimensionality of the equivalent elements for each element in the burst is 32. The similar reasoning to that used above leads to the completion of the proof for any burst of 1×2.

II4. Clearly, the proof for any rectangular burst errors of 2×1, 4×2, and 8×4 can be proved accordingly. Thus, we complete the proof of the Lemma for any rectangular bursts when $n = 3$.
II5. Now, we assume that the Lemma holds for any rectangular bursts when $n = N$. We are going to show that the Lemma holds for any rectangular bursts when $n = N + 1$. The same reasoning as used in I7, I8 and I9 can be applied to the current circumstances accordingly to complete the proof at this part.
II6. We therefore conclude the whole proof of the Lemma.

□

Now we are in a position to present the following theorem.

Theorem 2.1. *Consider a 2-D array of* $2^n \times 2^n$, *partition it into* 2^m *blocks for the integer* m *satisfying* $1 \leq m \leq 2n - 1$ *according to Definition 2.1. When* m *is even and* $m = 2k$, *any burst of* $2^k \times 2^k$ *in the interleaved array,* A, *obtained by using the successive packing is spread in the de-interleaved array such that each element of the burst falls into a distinct block of size* 2^{2n-2k}. *When* m *is odd and* $m = 2k + 1$, *any burst of* $2^k \times 2^{k+1}$ *or* $2^{k+1} \times 2^k$ *in* A *is spread in the de-interleaved array such that each element of the burst falls into a distinct block of size* $2^{2n-2k-1}$.

Meaning of Theorem 2.1. Theorem 2.1 states that in a 2-D array of $2^n \times 2^n$, A, interleaved with the successive packing technique, any square bursts of $2^k \times 2^k$ with $1 \leq k < n$ and any rectangular bursts of $2^k \times 2^{k+1}$ or $2^{k+1} \times 2^k$ with $0 \leq k < n$ can be spread such that each element in the burst falls into a distinct block in the de-interleaved array, where the block size, K, is 2^{2n-2k} for the burst of $2^k \times 2^k$, and $2^{2n-2k-1}$ for the bursts of $2^k \times 2^{k+1}$ or $2^{k+1} \times 2^k$. This indicates that, if a distinct code symbol is assigned to each element in the burst and all the code symbols associated with an individual K-equivalent class form a distinct codeword, then this technique guarantees that the burst error can be corrected with a one-random-error-correcting code. Furthermore, the interleaving degree equals to the size of the burst error, hence minimizing the number of codewords required in an interleaving scheme. In other words, with the successive packing technique the interleaving degree obtains the lower bound. In this sense, the successive packing interleaving technique is optimal. If a coding technique has a strong random-error-correcting capability, assume it can correct one error in every codeword of size four, then any burst errors of $2^{n-1} \times 2^{n-1}$ can be corrected. If a code, on the other hand, has less strong random-error-correcting capability, say, it can only correct one random error within a codeword of size 64, then with the successive packing only smaller burst errors, i.e., any burst of $2^{n-3} \times 2^{n-3}$ in the interleaved array can be corrected.

Proof. Due to Lemma 2.1, which establishes the equivalent-invariance with respect to burst translation for the types of burst errors under consideration, we only need to prove the theorem for the types of bursts located at the very north-west corner of the 2-D interleaved array, A. We prove the theorem for the square bursts first followed by a proof for the rectangular bursts. In either cases, the mathematical induction method is employed. Since in either case the proof is trivial as $n = 0, 1$, and is very simple as $n = 2$, we start the proof as $n = 3$ in either case.
A1. When $n = 3$, A is of 8×8, as shown in Figure 2 (a). $k < n$ implies that $k = 0, 1, 2$. Since the proof for $k = 0$ is trivial, we only prove the theorem for $k = 1, 2$.
A2. Consider $k = 2$. Look at the very northwest corner element, s_0. According to Definition 2.1, its 4-equivalent elements are: s_1, s_2, s_3, which are marked in Figure 2 and

are all outside of the square burst. Similarly, each element in the burst has its own 4-equivalent elements, which follow the same distribution pattern as the 4-equivalent elements of s_0. Hence all elements in the burst will be spread in the de-interleaved 2-D array so that there is only one code symbol, which is in error, within each single codeword of size 4. Consider $k = 1$. There are 4 elements in the northwest square burst. It can be seen from Figure 2 that all s_0's 16-equivalent elements are outside of the square burst. Similarly, each element in the burst follow the same distribution pattern as the 16-equivalent elements of s_0. Therefore, we can conclude that all elements in the burst will be spread into each single codeword of size 16 in the de-interleaved 2-D array. That is, the burst error can be corrected.

A3. After the proof for $n = 3$, we assume now that the theorem holds for any northwest square bursts in A when $n = N$. (Hence, A may be denoted by $S_{2^{2N}}$.) That is, as $n = N$, all the northwest square bursts of size $2^k \times 2^k$ with $k < N$ can be spread, where the corresponding codeword size becomes 2^{2N-2k}. To complete the proof of the theorem for all the northwest square bursts, according to the mathematical induction, we only need to show that the theorem holds for all the northwest square bursts in A when $n = N+1$. (Note that at this time A may be denoted by $S_{2^{2N+2}}$.)

A4. We prove the theorem first for $k < N$, then for $k = N$. Consider $k < N$. When $n = N$, s_0 has 2^{2N-2k} equivalent elements (including itself) in $S_{2^{2N}}$. When $S_{2^{2N+2}}$ is formed by using the successive packing with $S_{2^{2N}}$ according to Equation 2, the total number of the equivalent elements has increased by four times. Meanwhile, the difference of the subscripts of any two equivalent elements in the top-left quadrant of $s_{2^{2N+2}}$ has increased by four time compared with that in the $s_{2^{2N}}$. The increased equivalent elements are distributed evenly among the other three quadrants of $s_{2^{2N+2}}$ due to the successive packing (refer to Equation 2). It is easy to see that the equivalent elements of any specific elements in the northwest square burst of $2^k \times 2^k$ occupy the same positions within each of the four quadrants of $s_{2^{2N+2}}$. Furthermore, these positions are the same as that in the $S_{2^{2N}}$. It is hence observed that among the $2^{2N-2k+2}$ equivalent elements (including s_0) only s_0 is within the square burst of $2^k \times 2^k$ in $S_{2^{2N+2}}$.

Obviously, the above observation is also true for any elements in the northwest burst of $2^k \times 2^k$.

As mentioned above (in meaning of Theorem 2.1), all the $2^{2N-2k+2}$-equivalent elements form a codeword. The distribution pattern of all the $2^{2N-2k+2}$-equivalent elements discussed above indicates that there is only one error element in each codeword. Therefore, we finish the proof for $k < N$.

A5. When $k = N$, the total number of the equivalent elements of each element in the northwest square burst is 3. Hence the situation is similar to the square bursts of 4×4 when $n = 3$. We can use the reasoning, used in I2, to prove the theorem for $k = N$.

A6. Up to this point we have proved the theorem for all the northwest square bursts in A. The use of Lemma 2.1 at this point leads to the proof for all the square bursts in A.

A7. Now, prove the theorem for the rectangular bursts. Consider $n = 3$. All the six possible types of the rectangular bursts in this situation are: 1×2, 2×4, 4×8, 2×1, 4×2, and 8×4.

For the rectangular bursts of 4×8, located at the top-half of the 8×8 array. As shown in the proof of Lemma 2.1, the dimensionality of the equivalence is 2, and the equivalent element of each element in the burst is located in the bottom-half of A. We hence conclude that the burst can be corrected. For the northwest burst of 2×4 in A. At this time, the dimensionality of the equivalent class is 8. It is seen that within each of the 8-equivalent class there is only one element is in error (refer to Figure 2 (b)). Hence, the burst can be corrected. Clearly, the proof for the northwest rectangular burst errors of $1 \times 2, 2 \times 1, 4 \times 2$, and 8×4 can be proved accordingly. Thus, we complete the proof for all the northwest rectangular bursts when $n = 3$.

A8. Now, we assume that the theorem holds for any northwest rectangular bursts of $2^k \times 2^{k+1}$ or $2^{k+1} \times 2^k$ when $n = N$. Now prove the theorem for $n = N+1$. The reasoning used in A4 can be applied to the current circumstances accordingly to complete the proof when $k < N$.

A9. When $k = N$, the proof for the burst of $2^N \times 2^{N+1}$ can be conducted with the same reasoning as that used in proving the burst of 4×8 in the interleaved 2-D array of 8×8 (i.e., $n = 3, k = 2$). The proof for the burst of $2^{N+1} \times 2^N$ can be conducted accordingly.

A10. We therefore conclude by using the mathematical induction that any northwest rectangular bursts of $2^k \times 2^{k+1}$ or $2^{k+1} \times 2^k$ can be corrected. The use of Lemma 2.1 leads to that any burst errors of $2^k \times 2^{k+1}$ or $2^{k+1} \times 2^k$ in A can be corrected.

The proof of the theorem is thus completed.

□

3 Performance Comparison

We have compared the SP with the interleaving technique developed by the UPS (with the Boustrophedonic pattern) and that presented in [3]. Figure 3 shows the interleaved 2-D arrays of 8×8 obtained by applying these three tcchniques, respectively.

Firstly, look at the burst error of 2×2 located in the middle of the interleaved arrays. It involves the following four code symbols: $s_{20}, s_{62}, s_{43}, s_1$ in the interleaved array with the SP; $s_{35}, s_{27}, s_{28}, s_{36}$ with the UPS'; and $s_{35}, s_{43}, s_{60}, s_4$ with the method in [3]. After de-interleaving, the SP spreads the four elements in the burst error into four different quadrants, while either the UPS' method or the method in [3] has two error elements located within one quadrant. If a code can correct one random error in each codeword consisting of 16 code symbols, then only can the SP correct the 2×2 burst error.

Secondly, it is noticed that, when the interleaved 2-D arrays of 8×8 are considered, the SP can optimally correct the following types of bursts, i.e., any bursts of 2×4, or 4×2, or 2×2, according to Theorem 2.1; while the method in [3] can correct any bursts of size 4 with an interleaving degree of 8. Since any burst of size 4 in an array of 8×8 can be entirely included in either a burst of 2×2 or a burst of 2×4 or 4×2, the SP can also correct any bursts of size 4. As far as the efficiency is concerned, if a burst of size 4 can be included in a burst of 2×2, then as shown above, the SP is more efficient. If a burst of 4 is included in a burst of 2×4 or 4×2, then the interleaving degree is 8, indicating that the SP achieves the same efficiency in this case.

Thirdly, in addition to the three types of bursts described above, according to Theorem 2.1, the SP can optimally correct any bursts of 1×2, or 4×8, or 2×1, or 8×4, or 4×4 within an interleaved array of 8×8. It is hence observed that the SP is versatile.

4 Discussion

In addition to the 2-D arrays of $2^n \times 2^n$ (referred to as the case of base-2), the SP technique may extend to other kinds of 2-D arrays such as $3^n \times 3^n$ (the case of base-3), $5^n \times 5^n$ (the case of base 5), thus enhancing the SP's versatility. This is a subject under investigation.

Generalization of the SP technique to M-D ($M > 2$) array and applying the SP technique to M-D information protection are other subjects that need investigation.

Acknowledgments

Y. Q. Shi would like to acknowledge discussions with G. F. Elmasry and Y. G. Fang.

References

[1] S. B. Wicker, *Error Control System for Digital Communication and Storage*, Prentice-Hall, Inc, Englewood Cliffs, NJ, 1995.

[2] G. F. Elmasry and Y. Q. Shi, "An interleaving technique for error correction in 2-D bar coding," *Proceedings of Thirty-Sixth Annual Allerton Conference on Communication, Control, and Computing*, pp. 49–58, 1998.

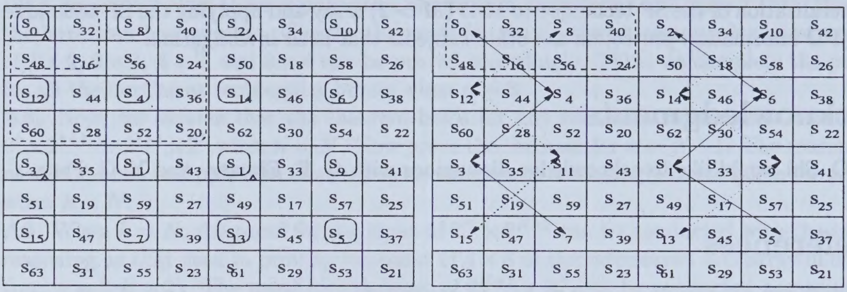
[3] M. Blaum, J. Bruck, and A. Vardy, "Interleaving schemes for multidimensional cluster errors," *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 730–743, 1998.

[4] K. A.S. Abdel-Ghaffar, "Achieving the Reiger Bound for Burst Errors Using Two-Dimensional Interleaving Schemes," *IEEE ISIT 1997*, p. 425, 1997.

[5] H. Imai, "Two-Dimensional Fire Codes," *IEEE Transactions on Information Theory*, vol. 19, no. 6, pp. 796–806, 1973.

[6] K. A.S.Abdel-Ghaffar, R. J.Mceliece, and Henk C.A.Van Tilborg, "Two-dimensional burst identification codes and their use in burst correction," *IEEE Transactions on Information Theory*, vol. 34, no. 3, pp. 494–504, 1988.

[7] Celso de Almeida and R. Palazzo Jr, "Two-dimensional interleaving using the set partitioning technique," *Proceedings. IEEE Int. Symp. Information Theory*, p. 505, 1994.



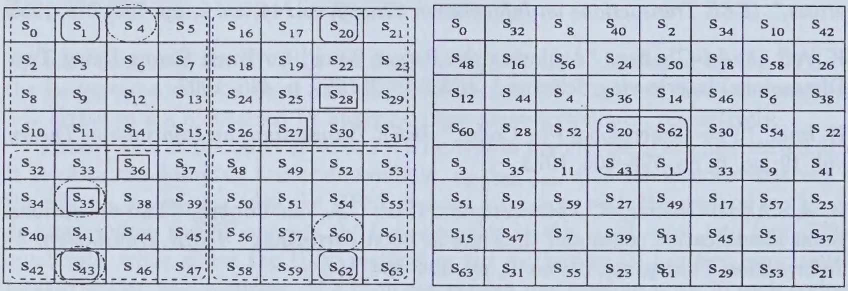
s_0	s_{32}	s_8	s_{40}	s_2	s_{34}	s_{10}	s_{42}
s_{48}	s_{16}	s_{56}	s_{24}	s_{50}	s_{18}	s_{58}	s_{26}
s_{12}	s_{44}	s_4	s_{36}	s_{14}	s_{46}	s_6	s_{38}
s_{60}	s_{28}	s_{52}	s_{20}	s_{62}	s_{30}	s_{54}	s_{22}
s_3	s_{35}	s_{11}	s_{43}	s_1	s_{33}	s_9	s_{41}
s_{51}	s_{19}	s_{59}	s_{27}	s_{49}	s_{17}	s_{57}	s_{25}
s_{15}	s_{47}	s_7	s_{39}	s_{13}	s_{45}	s_5	s_{37}
s_{63}	s_{31}	s_{55}	s_{23}	s_{61}	s_{29}	s_{53}	s_{21}

(a) The 4-($s_0 - s_3$) and 16-($s_0 - s_{15}$) equivalent element distribution patterns

s_0	s_{32}	s_8	s_{40}	s_2	s_{34}	s_{10}	s_{42}
s_{48}	s_{16}	s_{56}	s_{24}	s_{50}	s_{18}	s_{58}	s_{26}
s_{12}	s_{44}	s_4	s_{36}	s_{14}	s_{46}	s_6	s_{38}
s_{60}	s_{28}	s_{52}	s_{20}	s_{62}	s_{30}	s_{54}	s_{22}
s_3	s_{35}	s_{11}	s_{43}	s_1	s_{33}	s_9	s_{41}
s_{51}	s_{19}	s_{59}	s_{27}	s_{49}	s_{17}	s_{57}	s_{25}
s_{15}	s_{47}	s_7	s_{39}	s_{13}	s_{45}	s_5	s_{37}
s_{63}	s_{31}	s_{55}	s_{23}	s_{61}	s_{29}	s_{53}	s_{21}

(b) The 8-equivalent element distribution patterns

Figure 2: The square (a) and rectangular (b) equivalent element distribution patterns.

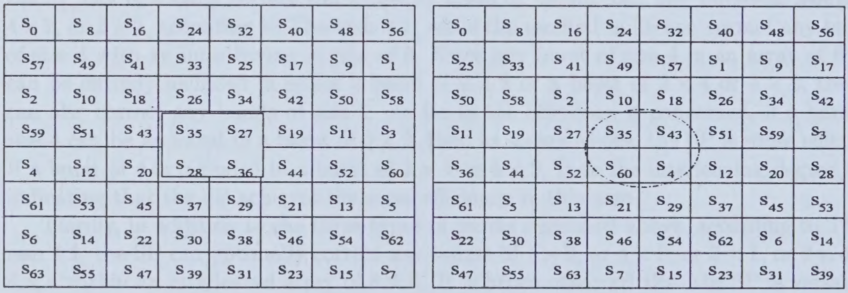


s_0	s_1	s_4	s_5	s_{16}	s_{17}	s_{20}	s_{21}
s_2	s_3	s_6	s_7	s_{18}	s_{19}	s_{22}	s_{23}
s_8	s_9	s_{12}	s_{13}	s_{24}	s_{25}	s_{28}	s_{29}
s_{10}	s_{11}	s_{14}	s_{15}	s_{26}	s_{27}	s_{30}	s_{31}
s_{32}	s_{33}	s_{36}	s_{37}	s_{48}	s_{49}	s_{52}	s_{53}
s_{34}	s_{35}	s_{38}	s_{39}	s_{50}	s_{51}	s_{54}	s_{55}
s_{40}	s_{41}	s_{44}	s_{45}	s_{56}	s_{57}	s_{60}	s_{61}
s_{42}	s_{43}	s_{46}	s_{47}	s_{58}	s_{59}	s_{62}	s_{63}

(a) De-interleaved array

s_0	s_{32}	s_8	s_{40}	s_2	s_{34}	s_{10}	s_{42}
s_{48}	s_{16}	s_{56}	s_{24}	s_{50}	s_{18}	s_{58}	s_{26}
s_{12}	s_{44}	s_4	s_{36}	s_{14}	s_{46}	s_6	s_{38}
s_{60}	s_{28}	s_{52}	s_{20}	s_{62}	s_{30}	s_{54}	s_{22}
s_3	s_{35}	s_{11}	s_{43}	s_1	s_{33}	s_9	s_{41}
s_{51}	s_{19}	s_{59}	s_{27}	s_{49}	s_{17}	s_{57}	s_{25}
s_{15}	s_{47}	s_7	s_{39}	s_{13}	s_{45}	s_5	s_{37}
s_{63}	s_{31}	s_{55}	s_{23}	s_{61}	s_{29}	s_{53}	s_{21}

(b) Interleaved array by the SP



s_0	s_8	s_{16}	s_{24}	s_{32}	s_{40}	s_{48}	s_{56}
s_{57}	s_{49}	s_{41}	s_{33}	s_{25}	s_{17}	s_9	s_1
s_2	s_{10}	s_{18}	s_{26}	s_{34}	s_{42}	s_{50}	s_{58}
s_{59}	s_{51}	s_{43}	s_{35}	s_{27}	s_{19}	s_{11}	s_3
s_4	s_{12}	s_{20}	s_{28}	s_{36}	s_{44}	s_{52}	s_{60}
s_{61}	s_{53}	s_{45}	s_{37}	s_{29}	s_{21}	s_{13}	s_5
s_6	s_{14}	s_{22}	s_{30}	s_{38}	s_{46}	s_{54}	s_{62}
s_{63}	s_{55}	s_{47}	s_{39}	s_{31}	s_{23}	s_{15}	s_7

(c) Interleaved array by using the Boustrophedonic pattern

s_0	s_8	s_{16}	s_{24}	s_{32}	s_{40}	s_{48}	s_{56}
s_{25}	s_{33}	s_{41}	s_{49}	s_{57}	s_1	s_9	s_{17}
s_{50}	s_{58}	s_2	s_{10}	s_{18}	s_{26}	s_{34}	s_{42}
s_{11}	s_{19}	s_{27}	s_{35}	s_{43}	s_{51}	s_{59}	s_3
s_{36}	s_{44}	s_{52}	s_{60}	s_4	s_{12}	s_{20}	s_{28}
s_{61}	s_5	s_{13}	s_{21}	s_{29}	s_{37}	s_{45}	s_{53}
s_{22}	s_{30}	s_{38}	s_{46}	s_{54}	s_{62}	s_6	s_{14}
s_{47}	s_{55}	s_{63}	s_7	s_{15}	s_{23}	s_{31}	s_{39}

(d) Interleaved array by Blaum et al's method

Figure 3: Performance comparison of three 2-D interleaving techniques.

Array Codes as Low-Density Parity-Check Codes

John L. Fan
Gigabit Wireless, Inc.
Mail: P.O. Box 221, Palo Alto, CA 94302-0221
E-Mail: jfan@dsl.stanford.edu

Abstract

Low-density parity-check (LDPC) codes are error-correcting codes defined by sparse parity-check matrices, and decoded using the message-passing algorithm. Array codes with an algebraic structure similar to Reed-Solomon codes are shown to possess sparse parity-check matrices, and hence are suitable for soft decoding as LDPC codes. These structured array codes are found to perform as well as random LDPC codes with equivalent parameters.

1 Array codes

The term "array codes" [1] refers to a class of algebraic error-correcting codes defined on arrays for use in detecting and correcting error bursts. We focus on the construction of array codes given in [2], which has an algebraic structure analogous to Reed-Solomon and BCH codes. These array codes possess an algebraic decoder that can correct one symbol error and multiple symbol erasures, making them suitable for correcting long bursts of erasures.

Low-density parity-check (LDPC) codes [4][5] are linear error-correcting codes defined by sparse parity-check matrices. The decoder is based on the message-passing algorithm, which involves passing probabilistic messages on a graph derived from the parity-check matrix, and can be used to achieve near-capacity performance. When viewed as a binary error-correcting code, the array code turns out to have a sparse parity-check matrix. This effort fits into the general context of finding algebraic constructions for LDPC and turbo-like codes. [6].

The array code is defined on a $p \times p$ binary array, where p is an odd prime. A parameter r ($\leq p$) specifies the number of parity-checks that the bits must satisfy. In general, an array $\{a_{i,j}\}$ is a codeword of the array code if it satisfies the following sets of constraints: for $0 \leq k \leq r-1$ and $0 \leq i \leq p-1$,

$$\bigoplus_j a_{\langle i+jk \rangle_p, j} = 0,$$

where the notation $\langle b \rangle_p$ means the unique number between 0 and $p-1$ inclusive such that $\langle b \rangle_p \equiv b \bmod p$. In addition, the last row is automatically set to be 0: $a_{p-1,j} = 0$ for $0 \leq j \leq p-1$. In other words, for $0 \leq k \leq r-1$, the elements on the slope-k diagonals sum to zero modulo 2, where the slope-k diagonal consists of the p elements that are obtained by moving downward k rows for each move to the column to the right, wrapping around the edges of the array. The first r columns contain the parity bits and the last $p-r$ columns contain the message bits.

2 Parity-check matrix

It is possible to gain flexibility in the block size and the rate through shortening the array code by removing some message symbols. Suppose that $p-n$ symbols are removed, leaving n symbols in the code that are indexed by

$$0 \leq \alpha_0 < \alpha_1 < \cdots < \alpha_{n-1} \leq p-1.$$

To construct the parity-check matrix for an array code, arrange the columns of the array into a single vector. Each vector v_i is a column vector of length p, with the last bit $v_{i,p-1}$ equal to zero. Let σ be the permutation matrix that consists of a single cyclic shift.

The punctured parity-check matrix then takes the following form:

$$\mathbf{H} = \begin{bmatrix} I & I & \cdots & I \\ \sigma^{\alpha_0} & \sigma^{\alpha_1} & \cdots & \sigma^{\alpha_{n-1}} \\ \vdots & \vdots & & \vdots \\ \sigma^{(r-1)\alpha_0} & \sigma^{(r-1)\alpha_1} & \cdots & \sigma^{(r-1)\alpha_{n-1}} \end{bmatrix}$$

The resulting block length is $n\,(p-1)$ bits (where the last bit of each p-block has been omitted), and the number of independent rows is $r\,(p-1)$, giving an overall rate of $\frac{n-r}{n}$.

In [3], it is shown that there are no cycles of length 4 for array codes, and criteria are given for choosing the columns so as to avoid other short cycles. In addition, simulations of an array code with parameters $p = 59$, $n = 30$, and $r = 3$ show that the array code performs the same as a random parity-check matrix with $N = 1740$ columns and $M = 174$ rows and $t = 3$ bits per column, where cycles of length 4 have been avoided. Hence, the added structure of the array code does not lead to a degradation in performance when compared to an equivalent random LDPC code.

In addition, this array code can concatenated with an inner error-correcting code to protect the symbols. This is especially appropriate if the decoder for the inner code can produce a symbol-wise erasure flag for use by the algebraic decoder for the array code.

References

[1] M. Blaum, P.G. Farrell, H.C.A. van Tilborg, "Array Codes," Chapter 22 in *Handbook of Coding Theory*, V.S. Pless and W.C. Huffman (Eds.), Elsevier Science B.V., 1998.

[2] M. Blaum, R. Roth, "New Array Codes for Multiple Phased Burst Correction," *IEEE Trans. Inform. Theory*, Vol. 39, No. 1, January 1993, pp. 66-77.

[3] J.L. Fan, "Array codes as LDPC codes," *Proc. Int'l. Symp. on Turbo Codes*, Brest, France, Sept. 2000, pp. 543-546.

[4] R.G. Gallager, "Low-Density Parity-Check Codes," *IRE Trans. Inform. Theory*, January 1962, pp. 21-28.

[5] D.J.C. MacKay, "Good Error-Correcting Codes Based on Very Sparse Matrices," *IEEE Trans. Inform. Theory*, vol. 45, no. 2, March 1999, pp. 399-431.

[6] R.M. Tanner, "Toward an Algebraic Theory for Turbo Codes," *Proc. Int'l. Symp. on Turbo Codes*, Brest, France, Sept. 2000, pp. 17–26.

Convolutional Coding Using Finite-Field Wavelets

F. Fekri, S. W. McLaughlin, R. M. Mersereau, R. W. Schafer

Dept. of Electrical and Computer Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0250
fekri@ee.gatech.edu

In earlier work [1, 2, 3, 4], we established a new framework to study error correcting codes. We developed the theory of finite-field wavelets. Next we applied this theory to the construction and decoding of block codes. Motivated by our previous results on wavelet-block coding, we proceed to study convolutional codes using finite-field wavelets. This method can be viewed as transmission of the message sequence over the wavelet basis functions. For convolutional coding we deal with infinite-length sequences. Thus, the filter bank representation of the wavelet systems does not perform cyclic operations as it did for block coding. We present some results on the algebraic properties of the wavelet encoders.

Let the set of the synthesis filters $\{g_0(n), \ldots, g_{L_1}(n)\}$ define an L-band orthogonal wavelet system. Then, a rate $1/L$ wavelet encoder can be constructed as shown in Fig. 1. The prefilters $\lambda_i(n)$ are introduced to keep the generality of the convolutional encoder.

Example 1. *Let consider the encoder shown in Fig. 1 constructed by a two-channel orthogonal filter bank in which* $g_0(n) = \{111011\}$ *and* $g_1(n) = \{110111\}$. *If we choose the prefilter* $\Lambda(z) = z^{-1}$, *the resulting convolutional code has constraint length 3 and* $d_{\text{free}} = 6$.

Generalization to arbitrary rate case is straightforward and we omit the detail here. An example of a rate $2/3$ wavelet convolutional encoder is given in Fig. 2. The encoder of the wavelet convolutional codes is the same as for wavelet block codes [4] except that the convolution, upsampling and downsampling operations are not circular. Furthermore, the syndrome generator and dual code of the wavelet convolutional codes can be obtained with the similar method we described for block codes [4]. In the following we state (without proof) some of the algebraic properties of the wavelet convolutional encoders.

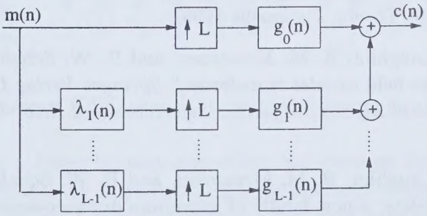


Figure 1: Filter bank structure of the rate $1/L$ wavelet encoder.

Proposition 1. *A rate* K/L *wavelet convolutional encoder is a basic encoder regardless of the choice of the impulse responses of the prefilters [5].*

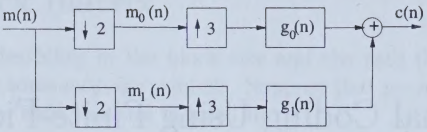


Figure 2: Example of a rate 2/3 wavelet convolutional encoder.

Corollary 1. *A rate K/L wavelet convolutional encoder is always non-catastrophic regardless of the choice of the impulse responses of the prefilters [5].*

We now address the following two questions. The first question is concerned with what we can say about wavelet convolutional encoders in terms of their number of abstract states. The second question is whether wavelet convolutional encoders are minimal in both the number of abstract states and the number of physical states.

Proposition 2. *For any arbitrary choice of the impulse responses of the prefilters, a rate K/L wavelet convolutional encoder matrix that is constructed by paraunitary filter banks contains a minimal number of abstract states over all equivalent encoding matrices for that code [5].*

Fact 1. *Any rate $1/L$ wavelet encoder of the form shown in Fig. 1 is minimal-basic regardless of the choice of the prefilters [5].*

Unfortunately, the above fact does not hold for arbitrary rate (K/L when $K > 1$) wavelet encoders in general.

It is important to notice two advantage of wavelet encoding systems with respect to conventional encoding methods. First, because of the orthogonality constraint, the wavelet search space for good convolutional codes is smaller than the traditional way of searching for the codes. Second, as shown by Corollary 1, the wavelet search method always leads to a non-catastrophic encoder.

References

[1] F. Fekri, R. M. Mersereau, and R. W. Schafer, "Realization of paraunitary filter banks over fields of characteristic two," in *Proc. Int. Conf. Acoust. Speech, and Signal proc.*, June 2000.

[2] F. Fekri, S. W. McLaughlin, R. M. Mersereau, and R. W. Schafer, "Error control coding using finite field wavelet transforms, Part II: Double circulant codes," *submitted to IEEE Trans. on Information Theory*, Decemebr 1999.

[3] F. Fekri, S. W. McLaughlin, R. M. Mersereau, and R. W. Schafer, "Double circulant self-dual codes using finite field wavelet transforms," *Springer Verlag Lecture Notes in Computer Science (LNCS);Applied Algebra, Algebraic Algorithms and Error-Correcting Codes*, pp. 355–364, 1999.

[4] F. Fekri, S. W. McLaughlin, R. M. Mersereau, and R. W. Schafer, "Rate $1/l$ block codes using finite field wavelets; a new family of maximum distance-separable codes and more," in *Proc. Conf. on Information Siences and Systems*, Princeton, NJ, March 2000.

[5] F. Fekri, S. W. McLaughlin, R. M. Mersereau, and R. W. Schafer, "Convolutional coding using finite field wavelet transforms: Time-varying codes and bipartite trellises," *in preparation.*

Low-Latency SISO via Message Passing on a Binary Tree [*]

Phunsak Thiennviboon and Keith M. Chugg
Dept. of Electrical Engineering-Systems
University of Southern California
Los Angeles, CA 90089-2565
{thiennvi,chugg}@usc.edu

Calculating the "soft inverse" of a *deterministic* finite-state machine (FSM) via the SISO (soft-in/soft-out) module is a key operation in many data detection/decoding algorithms [1, 3]. Perhaps the most appreciated application is iterative decoding of concatenated codes. The standard SISO algorithm is the forward-backward algorithm (e.g., [1, 3]) which can be derived from the generalized distributive law (GDL) algorithm [2] in a linear-structure junction tree as shown in Fig. 1(a). Note that a_k, s_k, t_k, and x_k are input, state, transition, and output at time index k of the FSM where $x_k = \mathcal{P}_k(s_k, a_k)$, $s_{k+1} = \mathcal{Q}_k(s_k, a_k)$, and $t_k = (a_k, s_k, s_{k+1}, x_k)$. The components in the dashed block define the SISO module. The local kernels of the nodes a_k and x_k are the soft-input for each variable and the local kernel of the node t_k is a *local* validity check for all variables of the node. The message passing schedule for this junction tree can be assigned by 3 familiar steps, i.e., forward recursion, backward recursion, and combining/completion. Note that the associated forward and backward recursion steps can be computed in parallel for all of the FSM states at a given time. This architecture provides $\mathcal{O}(N)$ latency and computational complexity, where N is the block size.

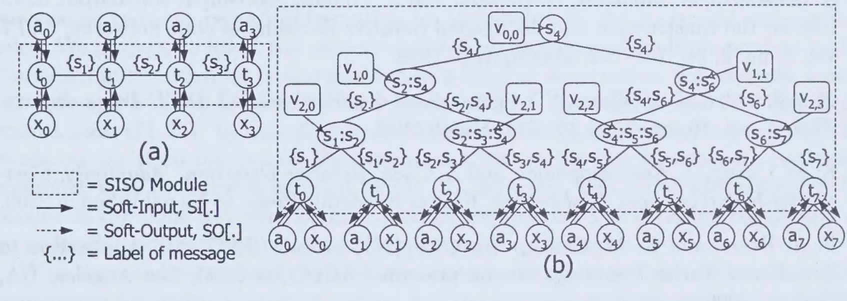

a0
a1
a2
a3
{s1}
{s2}
{s3}
t0
t1
t2
t3
x0
x1
x2
x3
(a)
= SISO Module
= Soft-Input, SI[.]
= Soft-Output, SO[.]
{...} = Label of message
V0,0
S4
{S4}
{S4}
V1,0
S2:S4
S4:S6
V1,1
V2,0
{S2}
{S2,S4}
V2,1
V2,2
{S4,S6}
{S6}
V2,3
S1:S2
S2:S3:S4
S4:S5:S6
S6:S7
{S1}
{S1,S2}
{S2,S3}
{S3,S4}
{S4,S5}
{S5,S6}
{S6,S7}
{S7}
t0
t1
t2
t3
t4
t5
t6
t7
a0
x0
a1
x1
a2
x2
a3
x3
a4
x4
a5
x5
a6
x6
a7
x7
(b)

Figure 1: Junction tree of SISO algorithm with (a) linear (b) binary-tree structure

In order to achieve a lower latency algorithm, we arrange the structure of the junction tree to be a binary tree by introducing additional local domains, $\{v_{i,j}\}$, as shown in Fig. 1(b) for $N = 8$. The local kernel of node $v_{i,j}$ is a *local* validity check for all variables of the node. Since both junction trees in Fig. 1 represent the same system, message passing (i.e., the GDL) on either produces the same desired soft inverse. Considering the activating schedule, all nodes at the same depth of the binary tree are activated simultaneously from the leaf nodes a_k and x_k to the root node $v_{0,0}$ (or inward recursion) , and then the process continues in the reverse direction (or outward recursion). Note that, the pending messages from nodes a_k and x_k to t_k are the soft-input, SI$[a_k]$ and SI$[x_k]$, and, after finishing outward recursion, the messages from node t_k to nodes a_k and x_k are the soft-output, SO$[a_k]$ and SO$[x_k]$ (see [3] for detailed description of the messages). We refer to this SISO as the *Forward-Backward-Tree SISO (FBT-SISO).* Using this inward/outward schedule, it is straightforward to show that the latency of this algorithm is $2\log_2 N$. Since the message updating on a node associated with 3 states typically has a highest complexity, the computational complexity of this algorithm is $\mathcal{O}(S^3 N)$ where S is maximum cardinality of all states in the FSM. Therefore, the FBT-SISO has $\mathcal{O}(\log_2 N)$ latency and $\mathcal{O}(N)$ computational complexity. Note that this binary-tree structure can be applied easily for an arbitrary value of N. As an example, let consider $N = 2^m$. The additional local domain $v_{i,j}$ ($i \in \{0, 1, ..., m-1\}$, $j \in \{0, 1, ..., 2^i - 1\}$) can be defined as follows: $v_{0,0} = s_{\alpha(0)}$, $v_{i\neq 0,0} = (s_{\alpha(i)}, s_{2\alpha(i)})$, $v_{i\neq 0, j\notin\{0,2^i-1\}} = (s_{(2j)\alpha(i)}, s_{(2j+1)\alpha(i)}, s_{(2j+2)\alpha(i)})$, and $v_{i\neq 0, 2^i-1} = (s_{(2^{i+1}-2)\alpha(i)}, s_{(2^{i+1}-1)\alpha(i)})$ where $\alpha(i) = 2^{m-i-1}$, i is the depth of node $v_{i,j}$ ($i = 0$ for root node $v_{0,0}$), j is the position (from left to right) of node $v_{i,j}$ at depth i.

---

[*] This work was supported in part by the National Science Foundation (NCR-CCR-9726391)

Recently, it has been demonstrated that the SISO computations can be done using a combination of prefix and suffix operations, which leads to a tree architecture with $\mathcal{O}(\log_2 N)$ latency [4]. This architecture is based on well-known tree-structures for fast parallel prefix computations in the Very Large Scale Integration (VLSI) literature (e.g., fast adders [5, 6, 7]), so we refer to it as a *tree-SISO.* As compared to the tree-SISO, the FBT-SISO presented herein has double the latency but is less (computationally) complex by a factor of approximately $\log_2 N$. Finally, there exist algorithms with a parallel prefix computation which have similar structure to the FBT-SISO, e.g., the fast adder in [7].

References

[1] S. Benedetto, G. Montorsi, D. Divsalar and F. Pollara, "Soft-input soft-output modules for the construction and distributed iterative decoding of code networks," *ETT* vol. 9, no. 2, pp. 155–172, March/April 1998.

[2] M. Aji, and R. J. McEliece, "The generalized distributive law," *IEEE Trans. Inform. Theory,* vol. 46, no. 2, pp.325-343, March 2000.

[3] K. M. Chugg, A. Anastasopoulos, and X. Chen, *Iterative Detection: Adaptivity, Complexity Reduction, and Applications.* Kluwer Academic Press, to be published in 2000.

[4] P. A. Beerel and K. M. Chugg, "An $\mathcal{O}(log(N))$-Latency SISO with Application to Broadband Turbo Decoding," to be presented MILCOM 2000, Los Angeles, CA, October 2000.

[5] A. E. Despain, "Chapter 2: Notes on computer architecture for high performance," in *New Computer Architectures*, Academic Press, 1984.

[6] R. P. Brent and H. T. Kung, "A regular layout for parallel adders," *IEEE Transactions on Computers*, vol. C-31, pp. 260–264, March 1982.

[7] T. H. Cormen, C. E. Leiserson and R. L. Rivest, *Introduction To Algorithms.* Cambridge, Mass.: The MIT Press, 1990.

Low Rate Product Codes*

Tingfang Ji, Wayne E. Stark

Dept. of Electrical Engineering and Computer Science, University of Michigan,
Rm. 4301 EECS Bldg, 1301 Beal Ave., Ann Arbor, MI 48109, USA.

Email: tji,stark@eecs.umich.edu

Iterative algorithms have been developed in literature for decoding product codes and provide superb performance [1, 2]. In a standard product code, the information bits form a $K \times K$ two-dimensional array, and each row and column of the array is encoded by an (N, K) block code. For each component code, a soft-input-soft-out decoder is utilized to estimate the likelihood of the information bits. After each iteration, the $2K$ component decoders exchange information, and use the estimate from other decoders as the *a priori* probability in the next iteration.

In this paper, we construct low complexity low rate product codes for iterative decoding. Most of the literature on iterative decoding of product codes has been focused on code rate greater than 0.6 due to complexity concerns. When trellis decoding is used to estimate the symbol likelihood of a block code, the decoding trellis usually grows exponentially with the number of parity check bits. Thus, high rate component codes are desired for constructing product codes. We also derive the weight enumerators for product codes with random interleavers and verify the analysis with computer simulations.

It is desirable to lower the rate of an existing product code without adding additional parity bits to the component code, since each additional parity check bit in the component code doubles the complexity of the trellis decoder. The original product codes are two-dimensional codes since each information bit is encoded by the horizontal and vertical codes. We propose to lower the code rate with high dimensional multiple product codes where each information bit is also encoded by a third or more codes. For example, a rate 0.63 code and a rate 0.57 code can be constructed from the original rate 0.72 Hamming product code $(31, 26)^2$ by encoding on the third and fourth dimension, respectively. To encode on the third dimension, the information bits in the 26×26 array are randomly scrambled, then each row is encoded by the $(31, 26)$ code. The fourth dimension can be encoded in the same way with another random interleaver. The decoding complexity of the multiple product codes increases linearly with the dimensions.

With the standard uniform random interleaver assumption, we can analyze the weight enumerators for multiple product codes. Let the component block code be an (N, K) linear code. Let the IOWE (input-output-weight-enumerator) of this block code be defined as $A^B(I, W) = \sum_{i=0}^{K} \sum_{w=0}^{N-K} A^B_{i,w} I^i W^w$, where $A^B_{i,w}$ is the number of codewords of input weight i and parity weight w. Consider an M-dimensional multiple product code with K^2 information bits. For each dimension, the component code, denoted by C, consists of K individual (N, K) block codes. The resulting IOWE, $A^C(I, W)$, is given by

---

*This work was supported in part by Raytheon Systems C. PO No. 7-620847, and DARPA GloMo Project, contract No. F30602-97-C-0314

$A^C(I,W) = (A^B(I,W))^K$. Using the random interleaver argument in [3], we can obtain the weight enumerator of the overall multiple product code from $A^C(I,W)$.

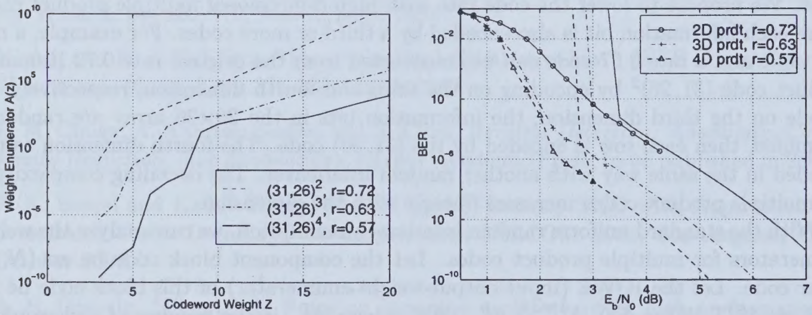
For high SNR (singal-to-noise ratio) the number of low weight codewords is important. In Fig. 1(a), the number of low weight codewords in 3D and 4D product codes are shown to be several orders of magnitude fewer than the original 2D product codes. The standard union bound is evaluated and the computer simulations are carried out to verify the performance of multiple product codes over AWGN channels. In Fig. 1(b), simulation results for BER for 2D, 3D, and 4D product codes are presented together with the union upper bounds. The union bounds are tight at hight SNR (signal to noise ratio), and diverge when SNR is below the cut-off rate threshold. For target BER of 10^{-3}, the required E_b/N_0 for the 4D product code is 1.5dB, which is 0.2dB and 0.8dB lower than the E_b/N_0 requirement for the 3D and 2D codes, respectively. For BER requirement of 10^{-8}, the gain increases to 0.8dB and 2.3dB, respectively. Compared with results of 2D product codes of similar rate and block size [1, 2], multiple product codes provide larger coding gains without significant complexity penalty.

References

[1] Joachim Hagenauer, Elke Offer, and Lutz Papke, "Iterative decoding of binary block and convolutional codes," *IEEE Transactions on Information Theory*, vol. 42, no. 2, pp. 429–445, Mar. 1996.

[2] Ramesh Mahendra Pyndiah, "Near-optimum decoding of product codes: block turbo codes," *IEEE Transactions on Communications*, vol. 46, no. 8, pp. 1003–1010, Aug. 1998.

[3] D. Divsalar, S. Dolinar, F. Pollara, and R. J. McEliece, "Transfer function bounds on the performance of turbo codes," Tech. Rep., JPL, Aug. 1995.



(a) Weight Enumerator for Low Weight Codewords

(b) Union Bounds and Simulations

Figure 1: Weight enumerators, union bounds and simulation results for multiple product codes(block length: 676 bits)

An Improved Iterative Decoding Algorithm for Turbo Product Code

Xuefei Wei, Ali N. Akansu
New Jersey Institute of Technology
Email: {xxw3114, ali}@oak.njit.edu

1 Introduction

Recently, great progresses have been made on the theoretical framework for iterative decoding[1, 2]. Both works indicate that the complexity of optimum decoding for concatenated code is prohibitively high. As a suboptimal algorithm, turbo decoding uses "divide-and-conquer" strategy: iteratively apply MAP decoding algorithm to component codes to approximate "global" optimum decoding. Intuitively, the first-activated constituent decoder contribute more in the decoding algorithm. For turbo product code, most erroneous blocks are due to this bias effect under high SNR . We propose to scale the extrinsic information before it is passed to successive decoder. This modified turbo decoding algorithm offers E_b/N_0 gain of about 0.35dB at BER of 10^{-5}.

2 Bias Effect

We discuss the decoding of two-dimensional $(n,k)^2$ product code, where each row and each column is a codeword of the (n,k) component code. Code set C^{-} contains codewords which satisfy the row constrains and code set $C^{|}$, the column constrains.

To find out how bias affect the final decoding result, we compare two decoding schedules. In schedule I, decoding begins from rows so that $C_1 = C^{-}$ and $C_2 = C^{|}$. In schedule II, $C_1 = C^{|}$ and $C_2 = C^{-}$. The AWGN-corrupted data blocks are classified into three types: convergent data blocks could be successfully decoded by either decoding schedules; double-error block could not be decoded by any of the two schedules; single-error blocks could be decoded by one of the two schedules. The presence of single-error blocks indicate the bias effect.

Simulations on two-dimensional $(16,11)$ extended Hamming product code are performed under different E_b/N_0. Optimum algorithm in [3] for block codes are used. For each value of E_b/N_0, 50,000 blocks are generated. For traditional iterative decoding under low E_b/N_0, most erroneous blocks belong to double-error type, which means they could not be decoded by either schedules. However, under high E_b/N_0, the two schedules only share a small portion of errorneous blocks. This implies that the bias effect in iterative decoding significantly degrades the performance at high SNR.

3 Scaled Factor Decoding

In traditional turbo decoding, decoder 2 takes channel values, together with extrinsic information $\mathbf{e}_1^{(1)}$ obtained from decoder 1, as a priori probability, though $\mathbf{e}_1^{(1)}$ is only a "local" optimization in the sense of MAP for C_1. Actually, the convergence step width provided by $\mathbf{e}_1^{(1)}$ might be too large so that decoder 2 fails to proceed toward the

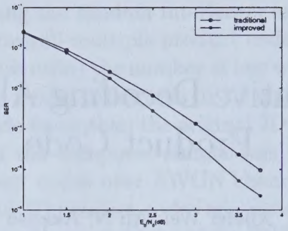


Figure 1: BER performance comparison

Table 1: Erroneous blocks under traditional and improved turbo decoding

	E_b/N_0	1	1.5	2	2.5	3	3.5
Traditional	double-error blk.	9813	3495	971	219	35	4
	single-error blk.	3205	2442	2219	490	133	31
Improved	double-error blk.	9861	3570	924	163	30	3
	single-error blk.	1066	585	222	55	9	2

correct contraction region. We propose to reduce the step width by scaling the extrinsic information before it is passed to the next decoder.

Let $\mathrm{e}_i^{(k)}$ denote the extrinsic information (probability) vector generated by decoder i at iteration k. We introduce a scaling factor $0.5 \leqslant \beta^{(k)} \leqslant 1$ to "shrink" $\mathrm{e}_i^{(k)}$ componentwise toward 0.5,

$$e'^{(k)}_{ij} = \beta^{(k)} e^{(k)}_{ij} + (1 - \beta^{(k)})(1 - e^{(k)}_{ij}) \tag{1}$$

where $\beta^{(k)}$ starts from a value in the range of (0.5, 1) and increases toward 1 with the iterations. Therefore, in the first several iterations, the "weight" of previous extrinsic information in the decoding process is reduced.

In the simulations, the evolution of scale factor $\beta^{(k)}$ with decoding iterations is set to

$$\beta^{(k)} = [0.75, 0.85, 0.95, 1, 1, \ldots]$$

As shown in Figure 1, the BER improvement increases as E_b/N_0 goes up. The SNR required to achieve a BER of 10^{-5} is about 0.35 dB less than that of traditional turbo decoder.

Under scaled factor decoding, bias effect still exists. However, single-error data blocks only constitute a small portion of error blocks. That is, the bias effect is much less severe when scaled factor is used. The total erroneous blocks are reduced, showing an improvement on block error rate.

References

[1] R. J. McEliece et.al, "Turbo decoding as an instance of pearl's belief propagation algorithm" in *IEEE J-SAC*, vol. 16, pp. 140-152, Feb. 1998

[2] M. Moher and T. A. Gulliver, "Cross-entropy and iterative decoding" in *IEEE Trans. IT*, pp. 3097-3104, Nov. 1998

[3] J. Hagenauer et.al , "Iterative decoding of binary block and convolutional codes" in *IEEE Trans. IT*, pp. 429-445, Mar. 1996

Encoder For Rotationally-Invariant Turbo TCM

Jing Sun, Oscar Y. Takeshita, and Michael P. Fitz
Department of Electrical Engineering
The Ohio State University
`sunj@ee.eng.ohio-state.edu; oyt@ee.eng.ohio-state.edu; fitz.7@osu.edu`

I. INTRODUCTION

After turbo code was introduced in 1993 [1], the same idea has been applied to Trellis Coded Modulation (TCM). Many Turbo TCM (TTCM) schemes have been proposed[2] [3]. Both high spectral efficiency and good performance can be achieved using these schemes.

A problem of coherent demodulation in TCM is phase ambiguity. Rotationally-Invariant (RI) TCM can be used to solve this problem. In an RI TCM, the encoder is designed so that all the rotated versions of a codeword are still codewords, and all these codewords map to exactly the same information sequence. Detailed study about RI TCM can be found in [4].

The RI idea can also be applied in TTCM design. Liu and Wilson [5] give the sufficient condition for a parallel concatenated TTCM code to be rotationally-invariant. In order for a TTCM to be RI, it is sufficient that the component codes are RI. However, as discovered for binary turbo code, the encoder structure is essential in the overall performance of the code. A good RI TCM encoder is not necessarily a good encoder for the RI TTCM.

II. ENCODER FOR RI TCM

In general, a TCM encoder can be regarded as a finite state machine (FSM).

$$s_k = f(s_{k-1}, i_k),\ c_k = g(s_{k-1}, i_k) \qquad (1)$$

where i, s, and c are input information symbol, state, and channel symbol, respectively. k is time index. Given an information sequence $\mathbf{i}$ and the initial state $s(0)$, the encoder will generate a code sequence $\mathbf{c}$ and a state sequence $\mathbf{s}$.

Define ρ_c to be the generating element of the rotation symmetry group of the constellation. Throughout this paper, we will use a QPSK constellation. Other constellations can be similarly considered. For QPSK, ρ_c can be the rotation by 90 degrees. Under ρ_c, the points in the constellation will form an orbit of size 4.

In order for an encoder to be rotationally invariant, it is sufficient that there exists a one to one correspondence ρ_s on state space, such that all the states of the encoder form orbits of size 4 under ρ_s, and ρ_s satisfies

$$\begin{array}{l} \rho_s(s_k) = f(\rho_s(s_{k-1}), i_k) \\ \rho_c(c_k) = g(\rho_s(s_{k-1}), i_k) \end{array} . \qquad (2)$$

In this way, applying ρ_c to a codeword $\mathbf{c}$ corresponds to encoding the same information sequence with initial state $\rho_s(s_0)$; We can use two variables to denote a state: $s = (ss, t_s)$. Superstate (ss) is the orbit the state belongs to. Type of state (t_s), distinguishes the different states in the orbit. It can take values from 0 to 3 along the orbit. We also define type of branch (t_b), which is the difference (mod 4) of the types of ending and starting states of a branch.

Using these notations, we can draw a simplified state transfer diagram (SSTD) for an RI TCM encoder. An example for an SSTD is given in Figure 1 for the RI TCM in [5]. In the SSTD, the branches are labeled with three terms. They are the input symbol, the branch type, and the output symbols corresponding to the accumulated branch types while traversing the

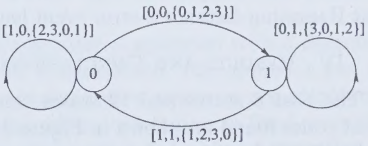


Fig. 1. SSTD for RI TCM encoder in [3]

This work was supported by NSF Grant ANI-9809018.

SSTD. By using this notation, we can see that the rotationally-invariant property of the encoder is natually implied. We can use this to enumerate possible RI TTCM encoders in our search.

III. ENCODER SEARCH FOR RI TTCM

In binary turbo codes, the distance between codewords corresponding to error events caused by input sequences that differ by Hamming distance 2 are important[6]. This is also true for TTCM. With this guideline, we hope to find the component RI TCM encoder with best input Hamming distance 2 spectrum. In particular, we want to maximize the effective free distance (the minimum Euclidean distance of all input Hamming distance 2 error events). If we have a tie in the previous step, then we maximize the second smallest effective distance, etc.

Unfortunately, this search involves a lot of computation. We use a suboptimum way to do this. First, we will ignore the output of the encoder and focus on maximizing the minimum length of the Hamming distance 2 error events. Then, we can properly assign outputs to the branches to maximize the effective free distance.

This method is valid most of the time since if the length of the minimum error event is longer, we have better chance to have larger effective free distance by properly assigning outputs to the branches. In order to cover the case that a shorter input Hamming distance 2 error events may have a larger Euclidean distance, we can relax the restriction of the first step to find schemes with comparably large shortest input Hamming distance 2 error event length.

IV. RESULTS AND CONCLUSIONS

The best 8 states and 12 states component codes found are shown in Figure 2 and 3 (in SSTD form):

Simulation results are shown in Figure 4 for rate 0.5bit/symbol RI TTCM schemes.

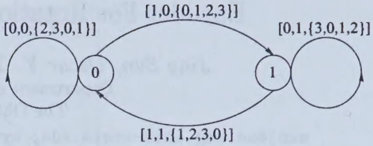


Fig. 2. SSTD of best 8 states code found

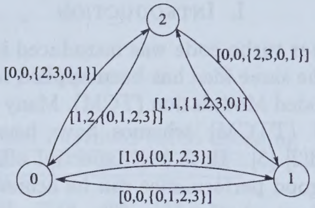


Fig. 3. SSTD of best 12 states code found

REFERENCES

[1] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near Shannon Limit Error-Correcting Coding and Decoding: Turbo-Codes," in *Proceedings of ICC'93*, Geneve, Switzerland, May 1993, pp. 1064–1070.

[2] P. Robertson and T. Wörz, "Bandwidth-Efficient Turbo Trellis-Coded Modulation Using Punctured Component Codes," *IEEE Journal on Selected Areas on Communication*, vol. 16, no. 2, pp. 206–218, February 1998.

[3] S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, "Parallel Concatenated Trellis Coded Modulation," in *IEEE International Conference on Communications*, 1996, vol. 2, pp. 974–978.

[4] M. D. Trott, S. Benedetto, R. Garello, and M. Mondin, "Rotational Invariance of Trellis Codes-Part I: Encoder and Precoders," *IEEE Transactions on Information Theory*, vol. 42, no. 3, pp. 751-765, May 1996.

[5] W. Liu and S. G. Wilson, "Rotationally-Invariant Concatenated (Turbo) TCM Codes," in *Conference Record of the Asilomar Conference on Signals, Systems and Computers*, May 1999, vol. 1, pp. 32–36.

[6] D. Divsalar and R. J. McEliece, "Effective Free Distance of Turbo Codes," *Electronics Letters*, vol. 32, no. 5, Feb. 1996.

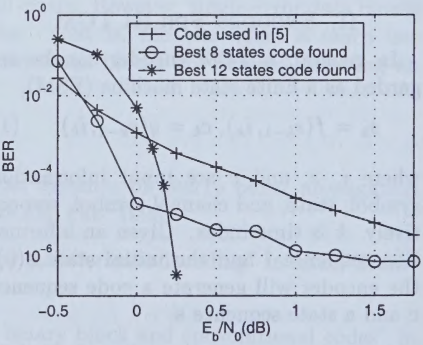


Fig. 4. Simulation results (4000 bits frame, 18 decoding iterations)

Efficient Implementation of a Maximum-Likelihood Detector for Space-Time Coded Systems[1]

M. Uysal and C.N. Georghiades
Electrical Engineering Department
Texas A&M University
College Station, Texas 77843-3128
E-mail: {georghia,uysal}@ee.tamu.edu

Abstract

It is well known that the quadratic form of the maximum-likelihood (ML) detector for space-time coded systems when channel state information (CSI) is not available does not readily lend itself to efficient implementation. In this paper we show that the generalized maximum-likelihood (GML) detector, whose structure coincides with that of the ML at high signal-to-noise ratios, can be expressed in a form that allows use of the Viterbi algorithm for efficient implementation. Simulation results over a mobile radio channel show the performance of the introduced receiver to be close to the optimum.

I. Introduction

A simple two-branch transmit diversity scheme has been introduced by Alamouti in [1] and has been shown to provide the same diversity order as maximal-ratio receiver combining (MRRC) with two receive antennas. Alamouti`s scheme is appealing in terms of its performance and simplicity. It results in a simple maximum-likelihood (ML) decoding algorithm based only on linear processing at the receiver. Space-time block coding, introduced in [2] and [3], generalizes Alamouti`s scheme to an arbitrary number of transmit antennas and is able to achieve the full diversity promised by the transmit and receive antennas. However, a main assumption in all these works is that channel state information (CSI) is perfectly available at the receiver. In this paper, we investigate efficient sequence estimation for space-time coded systems without assuming channel knowledge. We show that the generalized maximum-likelihood detector (GMLD) can be expressed in recursive form that allows the use of the Viterbi algorithm for efficient implementation. Previous work on sequence estimation for space-time coded systems in the absence of CSI using the expectation-maximization (EM) algorithm can be found in [4].

II. The ML Receiver

A wireless communication scenario, where the transmitter is equipped with L antennas and the receiver is equipped with a single antenna, is considered. The approach can be straightforwardly extended for the case when multiple receive antennas are employed. We will also focus on the quasi-static fading case where the fading is fixed over the received data block. Again, the more general case of non-static fading can be handled as well with a somewhat more involved notation.

---

[1] This work was partially supported by a gift from Texas Instruments.

Let the L fading coefficients corresponding to the L transmit-receive paths be denoted by α^i, $i = 1,2,\cdots,L$, and the transmitted data from the i-th transmit antenna at discrete time k be d_k^i. Then, the received data over a block of length N is

$$r = \underbrace{\begin{pmatrix} d_1^1 & d_1^2 & . & . & d_1^L \\ d_2^1 & d_2^2 & . & . & d_2^L \\ . & . & . & . & . \\ . & . & . & . & . \\ d_N^1 & d_N^2 & . & . & d_N^L \end{pmatrix}}_{D} \begin{pmatrix} \alpha^1 \\ \alpha^2 \\ \vdots \\ \alpha^L \end{pmatrix} + n, \tag{1}$$

where n is an independent Gaussian vector. Then, the GML detector is (see the Appendix)

$$\arg \max_{D} r^H D\left(D^H D\right)^{-1} D^H r\,. \tag{2}$$

Furthermore, for the Alamouti scheme, and orthogonal designs in general, we can write $D^H D = NEI_L$, where I_L is an identity matrix of dimension L. Therefore, the middle term in (2) is no longer sequence-dependent and can be neglected. Finally, we obtain the generalized maximum likelihood rule as,

$$\arg \max_{D} \left\| D^H r \right\|^2\,. \tag{3}$$

It can be shown that the GML receiver coincides with the true ML receiver for the Alamouti scheme, and orthogonal designs in general. It is clear from the form of the likelihood function in (3) that the it can not distinguish between a transmitted sequence and its negative. Thus it is necessary at the transmitter to limit the set of transmitted sequences of some length N to only half of the possible sequences, eliminating for every sequence included its negative. This can be accomplished by fixing the first symbol transmitted from each of the antennas to a specific value, i.e. inserting a pilot symbol at the beginning of each the frame.

III. Practical Implementation of the GML Receiver

Direct implementation of (3) is clearly impractical due to substantial computational complexity. However, it is possible to develop a recursive expression from (3). Expanding (3) and deleting terms not dependent on the data, we obtain

$$arg\ \max_{D}\ Re\left\{ \sum_{i=1}^{L} \sum_{k=2}^{N} r_k^* d_k^i \left[\sum_{l=1}^{k-1} r_{k-l} d_{k-l}^{i\,*} \right] \right\}, \tag{4}$$

where the inner term can be viewed as a channel estimate based on k-1symbols, i.e.

$$\hat{\alpha}_{k-1} = \sum_{t=1}^{k-1} r_{k-t} {d_{k-t}^{i}}^{*} \quad . \tag{5}$$

Instead of using all past data for channel estimation, we make use of only a fixed number and truncate the inner term corresponding to channel estimation. This approach is reasonable, since practical radio channels become uncorrelated in time. Even if the channel stays fixed for a long time, however, a reliable channel estimate can be obtained in this case by using only a truncated data block. Assuming that only J previous estimates are used, (4) may be rewritten as,

$$l(\boldsymbol{D}) = Re\left\{ \sum_{k=2}^{N} \left[\sum_{t=1}^{J} {r_k}^{*} r_{k-t} \sum_{i=1}^{L} d_k^i {d_{k-t}^i}^{*} \right] \right\} . \tag{6}$$

This leads to the recursive form,

$$l_N(\boldsymbol{D}) = l_{N-1}(\boldsymbol{D}) + Re\left\{ \sum_{t=1}^{J} {r_k}^{*} r_{k-t} \sum_{i=1}^{L} d_k^i {d_{k-t}^i}^{*} \right\} , \tag{7}$$

which can be implemented by a Viterbi-type algorithm.

IV. Performance

To demonstrate the applicability of recursive implementation, we consider Alamouti's scheme (i.e. the G_2 code in [2-3]) as an example. In Alamouti's scheme, at a given symbol period two signals are simultaneously transmitted from two antennas. If we denote the signal transmitted from Antenna 1 by s_1 and from Antenna 2 by s_2, during the next symbol period signal $-s_2^*$ is transmitted from Antenna 1 and signal s_1^* is transmitted from Antenna 2. Assuming truncation in (6) to two intervals, the trellis used in the computation of the branch metric is illustrated in Fig.1. We assign the message bits {00, 01, 10, 11} to trellis branches emanating from a specific node. Using (6), the accumulated branch metric for this code is given as,

$$\begin{aligned} Re\Bigg\{ \sum_{\substack{k=3 \\ k:odd}}^{N-1} \Big[& r_k^* r_{k-1} \left(-d_{k,1} d_{k-2,2} + d_{k,2} d_{k-2,1} \right) + r_k^* r_{k-2} \left(d_{k,1} d_{k-2,1}^* + d_{k,2} d_{k-2,2}^* \right) \\ & + r_{k+1}^* r_{k-1} \left(d_{k,2}^* d_{k-2,2} + d_{k,1}^* d_{k-2,1} \right) + r_{k+1}^* r_{k-2} \left(-d_{k,2}^* d_{k-2,1}^* + d_{k,1}^* d_{k-2,2}^* \right) \Big] \Bigg\} \end{aligned} \tag{8}$$

Fig. 2 shows simulation results for the benchmark performance of the optimum detector for both a quasi-static and a non-static fading channel. These results were obtained by an exhaustive search over all possible $\boldsymbol{D}$ matrices using the true likelihood functions (equation (3) for the quasi-static fading case and one for non-static fading channels not presented in this paper taken from [4]). Recalling that two symbol intervals are used to convey two symbols in Alamouti`s scheme, this search is limited to all possible 2^{N-2} $\boldsymbol{D}$ matrices. In order to keep the optimum decoder feasible, the frame size is chosen as 8 (i.e. N=8). For the quasi-static fading case, the channel is assumed to remain constant over a frame in the simulations. For the mobile radio channel, the channel is assumed to be constant during a symbol period, but to vary from symbol to symbol in a correlated fashion (Bessel function autocorrelation). At the start of each frame, the first two slots are used by pilot symbols. Therefore, for the frame size of *N*=8, the pilot insertion rate (PIR) is 1/4. The performance for various typical f_dT (normalized Doppler) values is illustrated in Fig. 2. It is observed that an error floor exists which is proportional to the channel f_dT value. The higher f_dT is, the higher the error floor. We also included the performance of Alamouti`s scheme with perfect channel state information (CSI) in Fig. 2 for comparison purposes. As in the original paper [1], for this case we assume that fading coefficients are constant over two symbols. This case can be viewed as a 'genie' bound where perfect knowledge of the channel is available, which is not a realistic assumption for a real channel [1].

In Figures 3-5, the performance of the recursive type receiver given by (8) over a mobile radio channel is demonstrated with the corresponding optimum decoder performance for f_dT values 0.05, 0.02 and 0.01, respectively. Also plotted in these figures, besides the ML detector's performance, is the performance of a receiver that first estimates the channel based on the pilot symbols and then uses it to detect the data symbols. The frame size is chosen again as *N*=8 in order to make the optimum decoder feasible and, therefore, the comparison between the two approaches possible. For all Doppler values, it is clearly seen that the performance of the proposed recursive receiver is very close to that of optimum one in a broad SNR region, and significantly outperforms the pilot-symbol assisted receiver, which has severe error floors, especially for large normalized Doppler values. The performance loss due to the recursive expression is less than 1 dB in most cases. We also observe that an error floor is observed in comparison to the optimum decoder. In the figures, we also included the recursive implementation with a larger frame size (i.e. *N*=20). The optimum decoder is not feasible in this case, therefore not included in the figures. As in all previous cases, two pilot symbols are used at the start of the frame, resulting in a PIR=1/10. In Fig. 3, we observe a performance loss as the frame size increases. This may be explained by the fact that we are operating at larger frame sizes than the Nyquist rate of the channel with an f_dT=0.05 value. In Figures 4-5, it is observed that the scheme with PIR=1/10 demonstrates slightly better performance than that with PIR=1/4. Although the small PIR value causes comparably poor channel estimates, the power loss due to PIR decreases, which results in an even slightly better performance compared to the scheme with frame size 4.

IV. Conclusion

We introduce a generalized ML detector which can be implemented using the Viterbi algorithm. Its performance is investigated over a mobile radio channel and shown to yield results close to the optimum over the same channel.

Appendix

In this appendix, we derive the generalized ML sequence estimator [5] and show that it is equivalent to ML sequence estimation for space-time block codes using the Alamouti scheme, or orthogonal designs in general.

Generalized ML sequence first finds an estimate for the channel, then uses this estimate in searching for the data matrix ***D***. The decoding rule is based on the maximization as follows,

$$\arg \max_{D} p(r|D,\hat{\alpha})$$
$$\hat{\alpha} = \arg \max_{\alpha} p(r|D,\alpha) \tag{9}$$

We have,

$$\ln p(r|D,\alpha) = \|r - D\alpha\|^2 = r^H D\alpha + \alpha^H D^H r - \alpha^H D^H D\alpha \tag{10}$$

Maximizing (10) w.r.t. α yields the channel estimate as,

$$\hat{\alpha} = \left(D^H D\right)^{-1} D^H r. \tag{11}$$

Inserting (11) in (10) and after some mathematical manipulation, we obtain

$$\arg \max_{D} \; r^H D\left(D^H D\right)^{-1} D^H r. \tag{12}$$

Recalling that $D^H D = NEI_L$ for space-time block codes, (12) reduces to

$$\arg \max_{D} \; \left\|D^H r\right\|^2 \tag{13}$$

which is equivalent to (3) (i.e. the ML receiver derived for the quasi-static case).

References

[1] S.M. Alamouti, "A Simple Transmit Diversity Technique for Wireless Communications", *IEEE Journal on Selected Areas in Communications*, Vol 16, No 8, pp.1451-1458, October 1998.

[2] V. Tarokh, H. Jafarkhani and A.R. Calderbank, "Space-Time Block Codes From Orthogonal Designs," *IEEE Transactions on Information Theory*, Vol 45, No 5, pp.1456-1467, July 1999.

[3] V. Tarokh, H. Jafarkhani and A.R. Calderbank, "Space-Time Block Codes For High Data Rate Wireless Communications: Performance Results," *IEEE Journal on Selected Areas in Communications*, Vol 17, No 3, pp. 451-460, March 1999.

[4] Y. Li, C.N. Georghiades and G. Huang, "Iterative Maximum Likelihood Sequence Estimation of Space-time Codes," submitted to the IEEE Transactions on Communications, July 1999.

[5] H.L. Van Trees, *Detection, Estimation and Modulation Theory*, New York: Wiley, 1968.

Fig.1. Trellis for the computation of recursive expression (Eq.8).

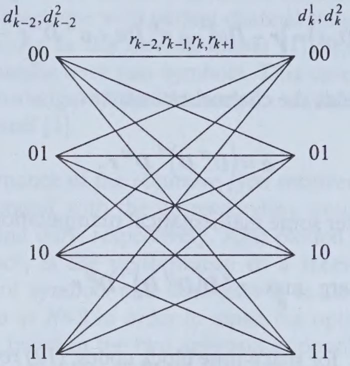

d_{k-2}^1, d_{k-2}^2
$r_{k-2}, r_{k-1}, r_k, r_{k+1}$
d_k^1, d_k^2
00
01
10
11
00
01
10
11

Fig.2. Optimum Decoder Performance (N=8)

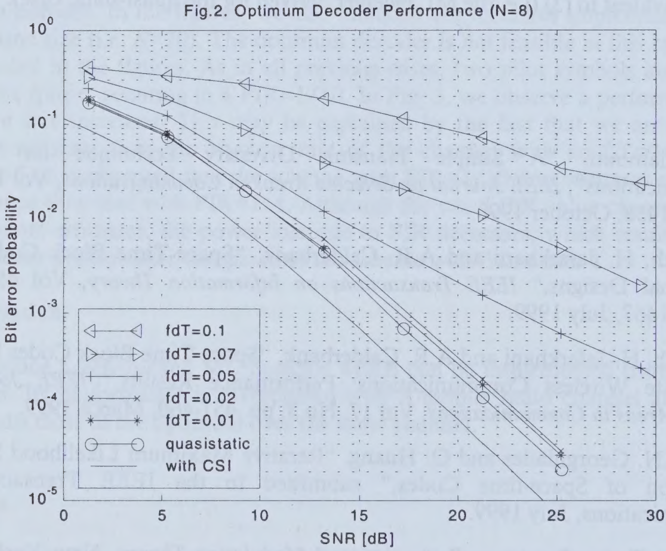


Fig.3. Recursive Decoder Performance (fdT=0.05)

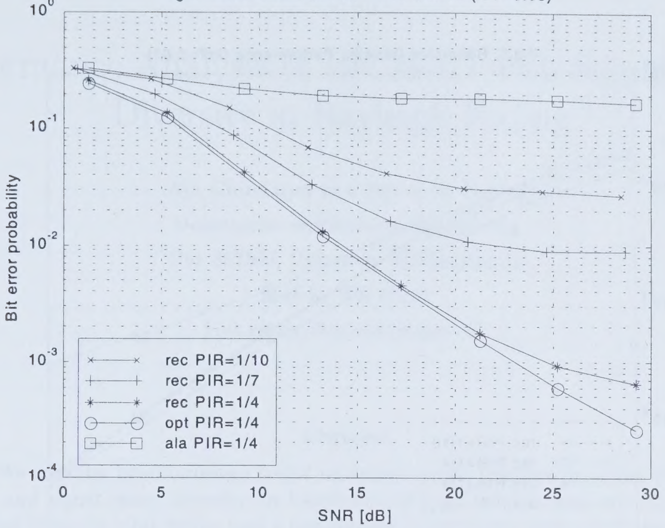
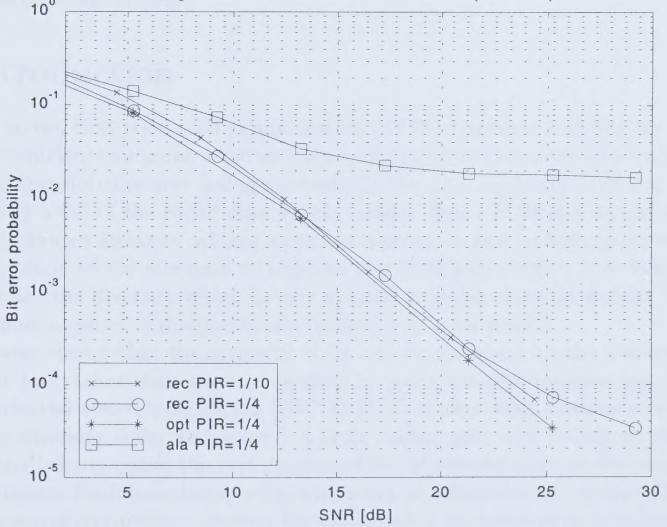


Fig.4. Recursive Decoder Performance (fdT=0.02)



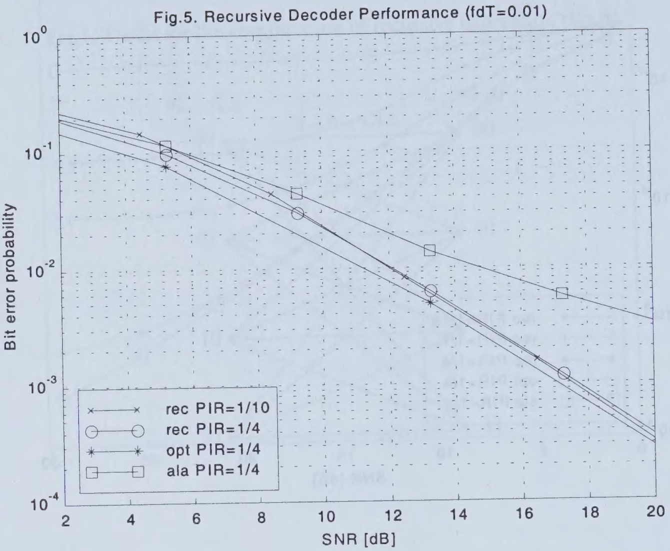


Fig.5. Recursive Decoder Performance (fdT=0.01)
Bit error probability
SNR [dB]
10^0
10^-1
10^-2
10^-3
10^-4
2
4
6
8
10
12
14
16
18
20
rec PIR=1/10
rec PIR=1/4
opt PIR=1/4
ala PIR=1/4

Performance Analysis of BICM-ID with Signal Space Diversity in Rayleigh Fading *

Aik Chindapol and James A. Ritcey
Department of Electrical Engineering
Box 352500, University of Washington
Seattle, WA 98195
{*aik,ritcey*}*@ee.washington.edu*

Abstract

We consider bit-interleaved coded modulation with iterative decoding (BICM-ID) and signal space diversity for bandwidth-efficient transmission over Rayleigh fading channels. Our design uses a non-rotated constellation with coordinate interleaving; hence, there is no constellation expansion as normally required for a signal space diversity technique. It is shown that diversity order can be improved to twice the diversity order of conventional BICM-ID; therefore, the code complexity can further be reduced while maintaining the same level of performance. Specifically, with the bandwidth efficiency of 2 bits/s/Hz over Rayleigh fading channels, a BER of 10^{-6} can be achieved with 4-state BICM-ID with signal space diversity at E_b/N_0 of about 7 dB. We also derive performance bounds for BICM-ID with signal space diversity over Rayleigh fading channels, which can be easily extended for other types of fading channels.

1 Introduction

It is well known that trellis-coded modulation (TCM) has been designed for bandwidth- and power-efficient communication over the additive white Gaussian noise (AWGN) channel. When transmitting over fading channels [1], the primary design criterion is shifted to the *diversity order* of the coded modulation system. Since TCM is designed to maximize the free Euclidean distance, adding a symbol interleaver and avoiding parallel transition are the common techniques used to improve the TCM performance over fading channels [1]. However, the diversity order for any symbol interleaved coded system is limited to the minimum number of distinct symbols along any error event.

It is later shown that the diversity order can be increased to the minimum number of distinct *bits* rather than *channel symbols* by using bit-wise interleaving. This allows a bit-interleaved coded modulation (BICM) [2, 3] system with moderate complexity to have large diversity order and achieve a large coding gain in a fading environment. In addition to diversity order, the performance of BICM also depends on the harmonic mean of the minimum Euclidean distance [3], which can be related to the product distance used in [1] as a secondary design criterion for trellis coded modulation over fading channels.

---

*This research is supported in part by ONR under grant N00014-99-1-0657.

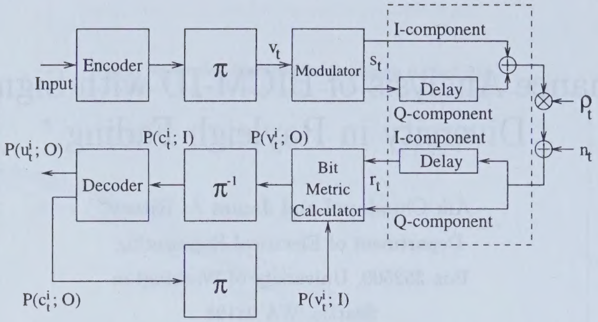


Figure 1: Block diagram of BICM-ID with the signal space diversity signal rotation. The delay implements coordinate interleaving. An increase in the complexity is shown in a dashed box.

An alternative scheme to increase diversity order involves interleaving over coordinates and choosing a proper signal constellation [4] and is known as *signal space* diversity [5]. The diversity is increased to the minimum number of distinct coordinates associated with any error event; therefore, this technique can provide significant improvement for uncoded [5, 6], TCM [7, 6] and BICM [8, 9] systems on Rayleigh fading channels. An attractive feature of signal space diversity over other schemes is that the performance over AWGN remains unchanged.

In [10, 11], Li and Ritcey show that BICM with iterative decoding (BICM-ID) greatly outperforms TCM and compare favorably with bandwidth-efficient turbo TCM [12] in AWGN. Since the design criterion depends on the channel model, we provide the design criteria and performance analysis of BICM-ID in Rayleigh fading channels in [13]. The main advantage of BICM-ID is that it requires only one soft-input soft-output decoder instead of two as normally used in turbo decoding [14, 12]. In [15], it is shown that BICM-ID can be added to turbo encoded BICM to achieve additional coding gain conventional system.

In this paper, we apply signal space diversity technique to BICM-ID. To minimize the system complexity, we propose a sub-optimal choice of using a non-rotated QAM constellation with coordinate interleaving. Our results show that the new combining scheme for BICM-ID significantly outperforms conventional BICM-ID in terms of the error floor performance.

Section 2 provides a brief discussion of BICM-ID with signal space diversity. An analytical bound for the probability of bit error and simulation results are shown in Section 3 and 4 respectively. Finally, Section 5 concludes the paper.

2 System Model

The block diagram for BICM-ID with signal space diversity is shown in Figure 1. At the transmitter, the information sequence is encoded by a convolutional encoder before being bit-wise interleaved. The purpose of the bit interleaver (π) is to break the sequential fading correlation and increased diversity order to the minimum Hamming distance of a code [2]. Next, m consecutive bits of the interleaved coded sequence are grouped to form $\boldsymbol{v}_t$, a channel symbol at the t^{th} signaling interval.

In order to achieve signal space diversity, the in-phase (I) and quadrature (Q) components of the same signal must be made to fade independently. This can be achieved by adding independent interleavers for I and Q components of a transmitted symbol. To maximize the diversity order, the constellation should be properly rotated such that all distinct signal symbols are separable on every coordinate [4]. However, an M-ary signal constellation is expanded to an M^2-ary constellation due to rotation and coordinate interleaving [9]. Therefore, in this paper we consider a sub-optimal but more practical choice of not expanding the signal set by using non-rotated QAM constellations with coordinate interleaving in order to minimize the system complexity.

Since the bit interleaver inherited by BICM already breaks the correlation due to fading, the purpose of coordinate interleaving is to make I and Q channels uncorrelated. Therefore, it need not be very deep if the channel has a short memory. However, coordinate interleaving implies that a QAM/PSK signal need to be quantized and stored which requires rapidly accessing and allocating large amounts of memory. We propose the practical solution of replacing coordinate interleavers with a time delay in only one of the quadrature components. With the fully-interleaved frequency non-selective fading assumption, a delay of at least a symbol period is required. However, a longer delay is needed if the channel has memory. We assume that the delay time exceeds the coherence time of the fading process. Note that the only increase in complexity compared to conventional BICM-ID are the delay components shown in the dashed box.

A modulator maps each channel symbol $\boldsymbol{v}_t$ to a complex signal $\boldsymbol{s}_t = \mu(\boldsymbol{v}_t) = s_{I_t} + js_{Q_t}$ chosen from a QAM constellation χ. We assume a frequency non-selective Rayleigh fading channel with coherent detection. With coordinate delay, the input to the bit metric calculator can be written as

$$\boldsymbol{r}_t = \rho_{t-k} s_{I_{t-k}} + j\rho_t s_{Q_{t-k}} + \boldsymbol{n}_t, \tag{1}$$

where k is the number of delayed symbols, ρ_t is the Rayleigh-distributed fading coefficient with $E(\rho_t^2) = 1$ and $\boldsymbol{n}_t$ is a complex white Gaussian noise sample with the variance of $N_0/2$. For the AWGN channel, $\rho_t = 1$. Throughout this paper, we assume perfect channel state information (CSI) so that ρ_t is perfectly estimated and available to the receiver.

With iterative decoding [11], the Viterbi decoder normally used in conventional BICM [2, 3] is replaced by the soft-input soft-output (SISO) decoder [16] and the output of the SISO decoder is fed back to the demodulator for bit metric recalculation. At the demodulator, the *a posteriori* probabilities for coded bits are calculated as

$$\begin{aligned} P(\boldsymbol{v}_t^i = b|\boldsymbol{r}_t) &\propto P(\boldsymbol{r}_t|\boldsymbol{v}_t^i = b)P(\boldsymbol{v}_t^i = b) \\ &= \sum_{\boldsymbol{s}_t \in \chi_b^i} P(\boldsymbol{r}_t|\boldsymbol{s}_t, \boldsymbol{v}_t^i = b)P(\boldsymbol{s}_t|\boldsymbol{v}_t^i = b)P(\boldsymbol{v}_t^i = b) \\ &= \sum_{\boldsymbol{s}_t \in \chi_b^i} P(\boldsymbol{r}_t|\boldsymbol{s}_t)P(\boldsymbol{s}_t), \qquad (2) \\ \text{where} \quad P(\boldsymbol{r}_t|\boldsymbol{s}_t) &= \frac{1}{\pi N_0}\exp\left(\frac{-(|\mathrm{Re}(\boldsymbol{r}_{t+k} - \rho_t \boldsymbol{s}_t)|^2 + |\mathrm{Im}(\boldsymbol{r}_{t+k} - \rho_{t+k}\boldsymbol{s}_t)|^2)}{N_0}\right), \end{aligned}$$

and χ_b^i is the subset of χ whose label has the binary value b at the i^{th} bit position.

The *a priori* probability $P(\boldsymbol{s}_t)$ is unavailable on the first pass of demodulation. Therefore, an equally-likely assumption is made. Following the notation of [16], we denote $P(q; I)$ as the *a priori* probability for a variable q. $P(q; O)$ is the *a posteriori* probability.

Note that $P(\boldsymbol{u}_t; I)$ is never available because it requires knowledge of the information sequence at the receiver.

On the second pass, the *extrinsic, a posteriori* probabilities $P(\boldsymbol{c}_t^i; O)$ put out by the SISO module are interleaved and fed back as the *a priori* probabilities $P(\boldsymbol{v}_t^i; I)$ to the demodulator. Therefore, *a priori* probabilities for (2) can be computed by

$$\begin{aligned} P(\boldsymbol{s}_t) &= P(\mu(\boldsymbol{v}_t; I)) \\ &= \prod_{i=1}^{m} P(\boldsymbol{v}_t^i = \tilde{\boldsymbol{v}}_t^i(\boldsymbol{s}_t); I), \end{aligned} \tag{3}$$

where $\tilde{\boldsymbol{v}}_t^i(\boldsymbol{s}_t)$ is the value of the i^{th} bit of the label corresponding to $\boldsymbol{s}_t = \mu(\tilde{\boldsymbol{v}}_t)$. Using (2) and (3), the *extrinsic a posteriori* bit probabilities for the second passes demodulation can be written as

$$\begin{aligned} P(\boldsymbol{v}_t^i = b; O) &= P(\boldsymbol{v}_t^i = b|\boldsymbol{r}_t)/P(\boldsymbol{v}_t^i = b; I) \\ &= \sum_{\boldsymbol{s}_t \in \chi_b^i} P(\boldsymbol{r}_t|\boldsymbol{s}_t) \prod_{\substack{j \neq i \\ j=1}}^{m} P(\boldsymbol{v}_t^j = \tilde{\boldsymbol{v}}_t^j(\boldsymbol{s}_t); I)). \end{aligned} \tag{4}$$

Equation (4) shows that we need only the *a priori* probabilities $P(\boldsymbol{v}_t^j; I)$ of the other bits $(j \neq i)$ on the same channel symbol $\boldsymbol{v}_t$ when recalculating the bit metrics. The receiver then uses (4) to *regenerate* the bit metrics and iterates demodulation and decoding. After the last pass, the final decoded outputs are the hard decisions based on the extrinsic bit probabilities $P(\boldsymbol{u}_t^i; O)$. This is the *total a posteriori* because $P(\boldsymbol{u}_t^i; I)$ is unused.

To reduce the implementation complexity, we use an additive "log-map" algorithm [16] for the SISO decoder and the log-sum is approximated by max operation aided by table lookups.

3 Analytical Bound

The union bound of probability of bit error for convolutional codes of rate k_c/n_c is given by [3]

$$P_b \leq \frac{1}{k_c} \sum_{d=d_{min}}^{\infty} W_I(d) f(d, \mu, \chi), \tag{5}$$

where $W_I(d)$ is the total input weight of error events at Hamming distance d and d_{min} is the minimum Hamming distance of the code. Note that $f(d, \mu, \chi)$ denotes the pairwise error probability (PEP) of BICM and depends only on Hamming distance d, a labeling map μ and a signal constellation χ.

From [3], the union bound of the PEP of BICM can be written in the form

$$\begin{aligned} f(d, \mu, \chi) &\leq \frac{1}{2\pi j} \int_{\alpha - j\infty}^{\alpha + j\infty} [\psi_{ub}(s)]^d \frac{ds}{s}, \\ \text{where } \psi_{ub}(s) &= \frac{1}{m 2^m} \sum_{i=1}^{m} \sum_{b=0}^{1} \sum_{\boldsymbol{x} \in \chi_b^i} \sum_{\boldsymbol{z} \in \chi_{\bar{b}}^i} \Phi_{\Delta(\boldsymbol{x}, \boldsymbol{z})}(s), \end{aligned} \tag{6}$$

and $\bar{b}$ is the complement of b. Note that $\Phi_{\Delta(\boldsymbol{x},\boldsymbol{z})}(s)$ is the Laplace transform of the probability density function of the metric difference $\Delta(\boldsymbol{x}, \boldsymbol{z})$ between the components $\boldsymbol{x}$

and $\boldsymbol{z}$. Due to large coding gain introduced by iterative decoding, we are most interested in an *analytical bound for the error floor*, to which the BICM-ID performance converges at very low BER.

For the analysis of the error floor of BICM-ID, we consider the idealized condition assuming error-free feedback. With perfect knowledge of *some* or *all* values of other bits forming a channel symbol, an M-ary constellation can be partitioned into sets of smaller constellations having larger inter-signal Euclidean distance. Note that this concept is similar to signal set partitioning used in trellis coded modulation [17]. Here, we choose to feedback the decoding decisions of all other bits; therefore, any M-ary signal set is converted to $M/2$ binary signal sets. This significantly increases the minimum inter-signal Euclidean distance and also reduces the number of nearest neighbor to one when Gray labeling is not used. Given ideal feedback for each $\boldsymbol{x} \in \chi_b^i$, $\chi_{\bar{b}}^i$ contains only one term $\tilde{\boldsymbol{z}} = \tilde{\boldsymbol{z}}(x)$, whose label has the same binary bit values as those of $\boldsymbol{x}$ except at the i^{th} bit position. Therefore, by removing the innermost summation in (6), the PEP of the error floor of BICM-ID can be written as

$$\begin{aligned} f(d,\mu,\chi) &\le \frac{1}{2\pi j}\int_{\alpha-j\infty}^{\alpha+j\infty}[\psi_{ef}(s)]^d\frac{ds}{s}, \qquad (7)\\ \text{where } \psi_{ef}(s) &= \frac{1}{m2^m}\sum_{i=1}^{m}\sum_{b=0}^{1}\sum_{\boldsymbol{x}\in\chi_b^i}\Phi_{\Delta(\boldsymbol{x},\tilde{\boldsymbol{z}})}(s). \end{aligned}$$

As pointed out in [3], the fading channel model used in conventional BICM does not apply if all transmitted signal components are affected by different fading coefficients. With coordinate interleaving, we assume that I and Q components of the same signal fade independently. Therefore, the equivalent received signal $\boldsymbol{r}$ after the transmission of the complex signal $\boldsymbol{s} = s_I + js_Q$ can be written as

$$\boldsymbol{r} = gs_I + jhs_Q + \boldsymbol{n}, \qquad (8)$$

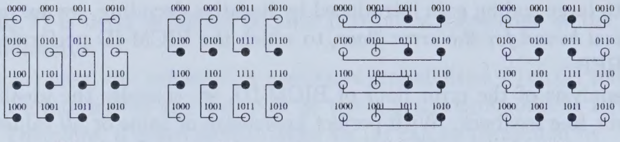
where g and h are the independent Rayleigh distributed fading gains, and $\boldsymbol{n} = n_I + jn_Q$ is complex white Gaussian noise with variance of $N_0/2$. Then, the conditional probability $p_\theta(\boldsymbol{r}|\boldsymbol{s})$ is given by

$$p_\theta(\boldsymbol{r}|\boldsymbol{s}) = \frac{1}{\pi N_0}\exp\left(\frac{-(|r_I - gs_I|^2 + |r_Q - hs_Q|^2)}{N_0}\right), \qquad (9)$$

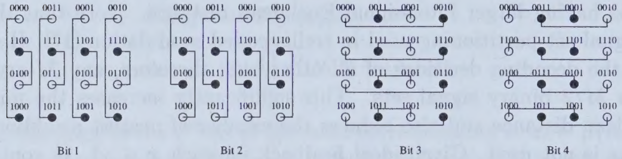
where θ is a vector containing values of g and h, and r_I and r_Q are the I and Q components of the received signal $\boldsymbol{r}$ respectively.

The metric difference between signals $\boldsymbol{x}$ and $\boldsymbol{z}$ when $\boldsymbol{y}$ is received can be written as

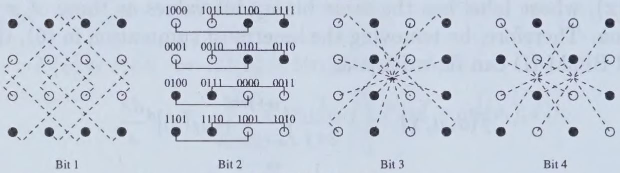
$$\begin{aligned} \Delta(\boldsymbol{x},\boldsymbol{z}) &= \log p(\boldsymbol{y}|\boldsymbol{x}) - \log p(\boldsymbol{y}|\boldsymbol{z})\\ &= g^2(x_I - z_I)^2 + 2g(x_I - z_I)n_I\\ &\quad + h^2(x_Q - z_Q)^2 + 2h(x_Q - z_Q)n_Q. \qquad (10)\\ \text{Then, } \Phi_{\Delta(\boldsymbol{x},\boldsymbol{z})}(s) &= E[e^{-s\Delta(\boldsymbol{x},\boldsymbol{z})}]\\ &= E_{g,h}\left[E[e^{-s\Delta(\boldsymbol{x},\boldsymbol{z})}|g,h]\right]\\ &= E_{g,h}\left[e^{g^2sd_I^2(N_0s-1)}e^{h^2sd_Q^2(N_0s-1)}\right].\\ &= E_{g,h}\left[e^{g^2s(x_I-z_I)^2(N_0s-1)}e^{h^2s(x_Q-z_Q)^2(N_0s-1)}\right]. \qquad (11) \end{aligned}$$


0000 0001 0011 0010
0100 0101 0111 0110
1100 1101 1111 1110
1000 1001 1011 1010
Bit 1
Bit 2
Bit 3
Bit 4

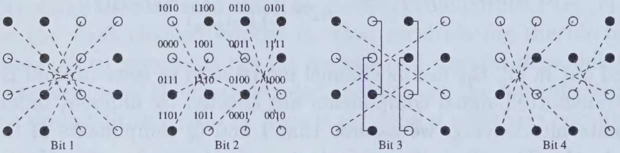
a. Gray Labeling


0000 0011 0001 0010
1100 1111 1101 1110
0100 0111 0101 0110
1000 1011 1001 1010
Bit 1
Bit 2
Bit 3
Bit 4

b. Mixed Labeling


1000 1011 1100 1111
0001 0010 0101 0110
0100 0111 0000 0011
1101 1110 1001 1010
Bit 1
Bit 2
Bit 3
Bit 4

c. Modified Set Partitioning Labeling


1010 1100 0110 0101
0000 1001 0011 1111
0111 1110 0100 1000
1101 1011 0001 0010
Bit 1
Bit 2
Bit 3
Bit 4

d. Random Labeling

Figure 2: 16QAM channel is converted to eight binary channels, each selected by the other three ideal feedback bits. The dashed lines show the binary signal sets that benefit from coordinate interleaving.

If g and h are independent and Rayleigh-distributed random variables, we have

$$\Phi_{\Delta(\boldsymbol{x},\boldsymbol{z})}(s) = \left(\frac{1}{1 + s(1 - sN_0) d^2_{I(\boldsymbol{x},\boldsymbol{z})}} \right) \left(\frac{1}{1 + s(1 - sN_0) d^2_{Q(\boldsymbol{x},\boldsymbol{z})}} \right), \tag{12}$$

where $d^2_{I(\boldsymbol{x},\boldsymbol{z})} = (x_I - z_I)^2$ and $d^2_{Q(\boldsymbol{x},\boldsymbol{z})} = (x_Q - z_Q)^2$. Note that (11) can be used to find the closed-form expression of $\Phi_{\Delta(\boldsymbol{x},\boldsymbol{z})}(s)$ for other channel models or coded modulation systems that utilize signal space diversity technique. Then, (7) can be used to evaluate the PEP when the Laplace transform of the probability density function of $\Delta(\boldsymbol{x}, \boldsymbol{z})(s)$ is defined in (12). The Gauss-Chebyshev quadrature [18, Sec. 13.3.2] method is used to numerically evaluate the PEP and finally the bit error probability of the error floor is obtained using (5).

Consider the impact of signal labeling on the performance of BICM-ID with signal space diversity. Since an M-ary constellation is converted to $M/2$ binary signal sets by iterative decoding, some of these binary signal sets can have independent I and Q components due to coordinate interleaving (delay) and still benefit from signal space diversity even though the constellation is not rotated. In specific, any binary signal set (given ideal feedback) lies vertically or horizontally in the signal space has either only I or Q component and does not gain additional diversity order as illustrated in Figure 2. It is obvious that Gray labeling and mixed labeling shown in Figure 2 do not benefit from signal space diversity without rotation because all of its binary signal sets lie vertically and horizontally. In other words, for Gray labeling and Mixed labeling, (12) can be rewritten as

$$\Phi_{\Delta(\boldsymbol{x},\boldsymbol{z})}(s) = \frac{1}{1 + s(1 - sN_0)\|\boldsymbol{x} - \boldsymbol{z}\|^2}, \tag{13}$$

which is $\Phi_{\Delta(\boldsymbol{x},\boldsymbol{z})}(s)$ for conventional BICM [3] and BICM-ID [9], for any pair of $\boldsymbol{x}$ and $\tilde{\boldsymbol{z}}$. Therefore, signal space diversity with a non-rotated constellation has no effect on the probability of bit error of the error floor of BICM-ID for these two labeling maps.

This suggests that Modified Set Partitioning (MSP) labeling and Random labeling are suitable for BICM-ID with signal space diversity when a constellation is not rotated. Note that MSP labeling and Random labeling are specifically designed for BICM-ID over fading channels according to the design guidelines shown in [13].

4 Simulation Results

In this section, we show simulation results for BICM-ID with signal space diversity (SSD) over the Rayleigh fading channel. We limit our attention to transmitting the information of 2 bit/s/Hz. Therefore, a 16QAM constellation is used with a rate 1/2 convolutional codes with maximal $d_{\min}$. Unless it is stated otherwise, each data block contains 2000 information bits and the spread-random interleavers described in [10] are used. The error floor (EF) bounds are compared with the simulated results to illustrate the closeness of the bound. For each simulated value, more than 10 errors are observed.

Figure 3 and 4 show the performance of 4-state BICM-ID with signal space diversity using MSP and Random labeling maps respectively. Due to an increase in diversity order, it can be noticed that the slope of the error floor of BICM-ID with SSD is steeper than that of conventional BICM-ID due to an increase in diversity order. Therefore, the error floor is significantly lowered by more than one order of magnitude for both labeling maps. This allows BICM-ID with MSP labeling to reach BER of 10^{-6} at E_b/N_0 of 7 dB with only 6 passes. This performance is comparable to many turbo coded modulation operated at the same spectral efficiency of 2 bits/s/Hz. For example, a 16-state turbo code shown in [14] reaches BER of 10^{-5} with 4 passes at E_b/N_0 of 6 dB and an 8-state parallel concatenated TCM (PCTCM) achieves BER of 10^{-6} with 5 passes at E_b/N_0 of 7 dB. It should be noted that these turbo coding schemes require *two* SISO decoders with larger complexity than a *single* 4-state SISO decoder used in BICM-ID.

In Figure 5, we compare the performance of 16-state BICM-ID and 4-state BICM-ID with SSD. The MSP labeled 16QAM constellation with the rate 1/2 convolutional codes are used with both BICM-ID schemes. Gray-labeled 16QAM BICM with the *de facto* standard 64-state, rate 1/2 code [3] is also shown for comparison. It can be seen that after 6 passes, BICM-ID with SSD can achieve almost the same performance as conventional BICM-ID having *four* times a number of decoding states and significantly outperforms conventional 64-state BICM.

5 Conclusions

In this paper, we analyze the performance of bit-interleaved coded modulation with iterative decoding (BICM-ID) and signal space diversity. It is shown that signal space diversity can increase diversity order, the primary design criterion for digital communication over fading channels; therefore, it can significantly improve the performance of BICM-ID in Rayleigh fading. We propose a sub-optimal design that utilizes the carefully designed labeling maps and coordinate delay in order to minimize the system complexity. Therefore, the code complexity can be significantly reduced while maintaining about the same level of performance and complexity. An analytical bound for the error floor of BICM-ID with signal space diversity is also derived and verified with the simulation results.

References

[1] D. Divsalar and M. Simon, "The design of trellis coded MPSK for fading channel: Performance criteria," *IEEE Trans. Commun.*, vol. 36, pp. 1004–1012, Sept. 1988.

[2] E. Zehavi, "8-PSK trellis codes for a Rayleigh channel," *IEEE Trans. Commun.*, vol. 40, no. 5, pp. 873–884, May 1992.

[3] G. Caire, G. Taricco, and E. Biglieri, "Bit-interleaved coded modulation," *IEEE Trans. Inform. Theory*, vol. 44, no. 3, pp. 927–945, May 1998.

[4] K. Boulle and J. C. Belfiore, "Modulation scheme designed for the Rayleigh fading channel," in *Proc. of CISS'92*, March 1992, pp. 56–67.

[5] J. Boutros and E. Viterbo, "Signal space diversity: A power- and bandwidth-efficient diversity technique for the Rayleigh fading channel," *IEEE Trans. Inform. Theory*, vol. 44, no. 4, pp. 1453–1467, July 1998.

[6] S. B. Slimane, "An improved PSK scheme for fading channels," *IEEE Trans. Veh. Technol.*, vol. 47, no. 2, pp. 703–710, May 1998.

[7] B. D. Jelicic and S. Roy, "Design of a trellis coded QAM for flat fading and AWGN channel," *IEEE Trans. Veh. Technol.*, vol. 44, pp. 192–201, Feb. 1995.

[8] D. L. Goeckel, "Coded modulation with non-standard signal sets for wireless OFDM systems," in *Proc. of ICC'99*, June 1999, pp. 791–795.

[9] A. Chindapol and J. A. Ritcey, "BICM with signal space diversity: Bounds for rotated BICM and BICM-ID," in *Proc. of CISS'00*, Mar. 2000.

[10] X. Li and J. A. Ritcey, "Trellis coded modulation with bit interleaving and iterative decoding," *IEEE J. Select. Areas Commun.*, vol. 17, no. 4, pp. 715–724, Apr. 1999.

[11] X. Li and J. A. Ritcey, "Bit interleaved coded modulation with iterative decoding," in *Proc. of ICC'99*, June 1999, pp. 858–864.

[12] P. Robertson and T. Worz, "Bandwidth-efficient turbo trellis-coded mudulation using punctured component code," *IEEE J. Select. Areas Commun.*, vol. 16, pp. 206–218, Feb. 1998.

[13] A. Chindapol, J. A. Ritcey, and X. Li, "BICM-ID for QAM constellations on Rayleigh fading channels," in *Proc. of CISS'99*, Mar. 1999, pp. 57–61.

[14] S. L. Goff, A. Glavieux, and C. Berrou, "Turbo-codes and high spectral efficiency modulation," in *Proc. of ICC'94*, May 1994.

[15] I. Abramovici and S. Shamai, "On turbo encoded BICM," *Annales des Telecommunications*, vol. 54, no. 3-4, pp. 225–234, March-April 1999.

[16] S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara, "A soft-input soft-output APP module for iterative decoding of concatenated codes," *IEEE Communications Letters*, vol. 1, no. 1, pp. 22–24, Jan. 1997.

[17] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Trans. Inform. Theory*, vol. IT-28, pp. 56–67, Jan. 1982.

[18] S. Benedetto and E. Biglieri, *Principles of Digital Transmission with Wireless Applications*, New York, NY: Kluwer, 1999.

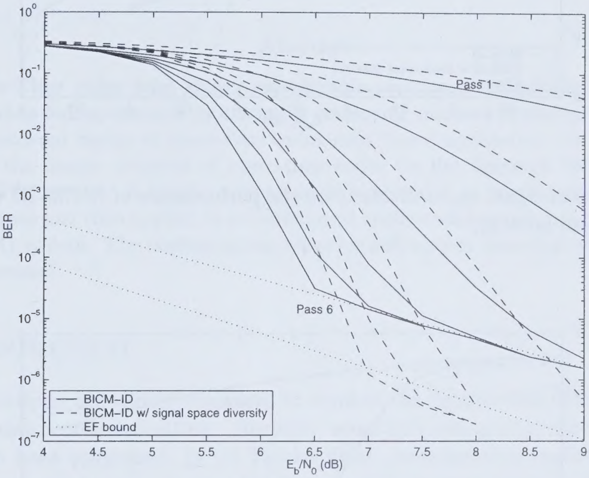


Figure 3: Effects of signal space diversity on the performance of BICM-ID with a 4-state code and MSP labeling.

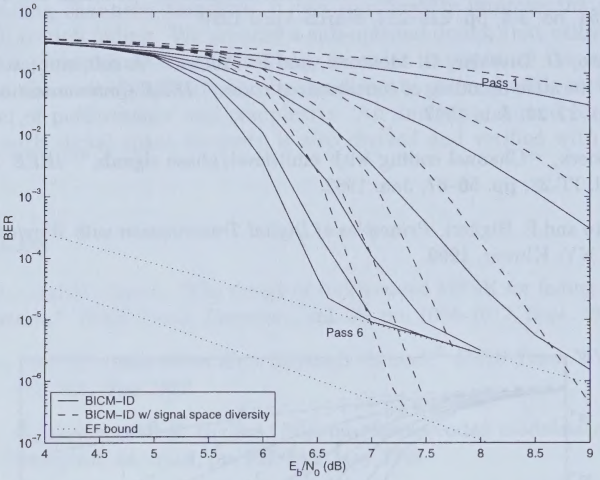


Figure 4: Effects of signal space diversity on the performance of BICM-ID with a 4-state code and Random labeling.

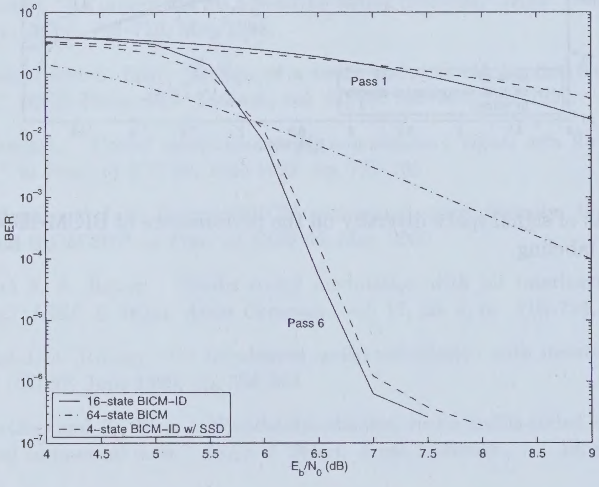


Figure 5: Performance comparison between 16-state BICM-ID, 64-state BICM and 4-state BICM-ID with signal space diversity (SSD) in Rayleigh fading. A 16-QAM constellation is used with rate 1/2 convolutional codes in all cases. MSP labeling and Gray labeling are used with both BICM-ID schemes and conventional BICM respectively.

Space-Time Coding for Rayleigh Fading Channels in CPM system

Xiaoxia Zhang
Dept. of Electrical Engineering
The Ohio State University
Columbus, OH 43210-1272
xiaoxia@ee.eng.ohio-state.edu

Michael Fitz
Dept. of Electrical Engineering
The Ohio State University
Columbus, OH 43210-1272
fitz@ee.eng.ohio-state.edu

Abstract

Space-Time codes have shown considerable promise for reliable transmission over wireless fading channels by efficiently employing transmit diversity. Past work only considered design of space-time codes with linear modulation. In this paper we give the design criterion of space-time codes for flat Rayleigh Fading channel in CPM system. A simple code example is provided to demonstrate its good performance and then applied to an interleaved trellis-coded space-time CPM (TC-ST-CPM) system. The corresponding adaptive soft-output detection algorithm is also presented.

1 Introduction

Diversity techniques are widely employed to combat the detrimental effects of multipath fading in wireless communications. Recently some interesting approaches for transmit diversity have been suggested. In [1] Tarokh *et al.* introduced space-time trellis coded modulation proposing a joint design of coding, modulation and transmit diversity for fading channels. It can achieve the same diversity advantage as receive diversity by avoiding destructive superposition of the signals transmitted simultaneously from different antennas. Meanwhile, for the transmission of digital information over both bandwidth- and power-limited channels such as mobile satellite communications and land mobile radio communications, continuous phase modulation (CPM) is an attractive modulation scheme [2]. It is worth investigating the design of codes making use of transmit diversity in CPM system. This paper presents the design criterion for designing space-time codes in CPM system. A code example to achieve maximum diversity level is given. This code is then used in an interleaved TC-CPM scheme to obtain better performance. The corresponding adaptive soft-output detection algorithm without perfect channel state information is also provided.

2 System Model

We consider a mobile communication system where there are L_t transmit antennas and L_r transmit antennas. In the transmitter shown in Fig. 1, data is encoded by the channel encoder, the encoded data goes through an interleaver and passes the space-time encoder

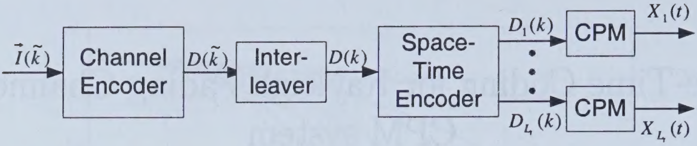


Figure 1: Transmitter Block Diagram

to get L_t streams of data. Each stream of data is used as the input to a CPM modulator. The modulated signals are transmitted through L_t transmit antennas at the same time and the signal received at each receive antenna is a noisy superposition of the L_t transmitted signals corrupted by Rayleigh fading and an independent zero-mean complex AWGN. If we limit the CPM scheme to be of single modulation index, the signal received by antenna j, $Y^j(t)$, can be given by [2]:

$$Y_j(t, \mathbf{D}) = \sum_{i=1}^{L_t} C_{i,j}(t) X_i(t, \vec{D}_i) + N_j(t), \quad j = 1, \cdots, L_r \tag{1}$$

$$X_i(t, \vec{D}_i) = \sqrt{\frac{E_s}{T}} \exp\left(j\phi(t, \vec{D}_i)\right), \quad i = 1, \cdots, L_t \tag{2}$$

$$\phi(t, \vec{D}_i) = 2\pi h_i \sum_{l=0}^{k} D_i(l) q_m(t - lT), \quad (k-1)T \le t < kT \tag{3}$$

where $D_i(k)$ is the kth modulation symbol of the ith antenna, h_i is the modulation index, $q_m(t)$ is the phase smoothing response function, T is the symbol time and E_s is the symbol energy. It can be shown that for $k \ge L_m$ where L_m is the modulation memory [2]

$$\phi(t, \vec{D}_i) = \Theta_i(k) + 2\pi h_i \sum_{l=k-L_m+1}^{k} D_i(l) q(t - lT) \tag{4}$$

If the modulation index h_i is rational, say $h_i = \dfrac{2m}{p}$ where m and p are relatively prime integers, then $\Theta_i(k) \pmod{2\pi} \in \Omega_\Theta$, where [2]

$$\Omega_{\Theta_i} = \left\{ 0, \frac{2\pi}{p}, \frac{4\pi}{p} \cdots \frac{2(p-1)\pi}{p} \right\}. \tag{5}$$

For rational modulation indices $\phi(t, \vec{D}_i(k))$ is constrained to lie on a trellis. The state of the system $\vec{S}_i(k)$ at time kT, is defined as

$$\vec{S}_i(k) = [\, \Theta_i(k) \;\; D_i(k - L_m + 1) \;\; \cdots \;\; D_i(k-1) \,] \tag{6}$$

With these definitions, $[D_i(k), \vec{S}_i(k)]$ uniquely specify $\phi(t, D_i(k), \vec{S}_i(k))$ and $X^i(t, \vec{D}_i)$ in the semi-open interval $[kT, (k+1)T)$.

3 Code Design

3.1 Design Criterion

Assume a transmission of length N_c, the maximum likelihood metric of sequence $\vec{\mathbf{D}} = \vec{\alpha}$ in the presence of perfect channel state information $\mathbf{c}(t)$ is:

$$m(\mathbf{y}(t)|\mathbf{c}(t), \mathbf{x}(t, \vec{\alpha})) = -\sum_{j=1}^{L_r} \int_0^{N_c T} |y(t) - \sum_{i=1}^{L_t} c_{i,j}(t) x_i(t, \vec{\alpha}_i)|^2 dt. \tag{7}$$

The pairwise error probability (PWEP) of confusing $\vec{\mathbf{D}} = \vec{\alpha}$ with another sequence $\vec{\mathbf{D}} = \vec{\beta}$, denoted as $P(\vec{\alpha} \to \vec{\beta})$, is given by:

$$P(\vec{\alpha} \to \vec{\beta}) = \Pr\left\{ m(\mathbf{y}(t)|\mathbf{c}(t), \mathbf{x}(t, \vec{\beta})) \geq m(\mathbf{y}(t)|\mathbf{c}(t), \mathbf{x}(t, \vec{\alpha})) \right\} \tag{8}$$

Following the similar method in [3], it can be shown that the pairwise error probability can be upper bounded by:

$$P(\vec{\alpha} \to \vec{\beta}) \leq \left[\frac{\binom{2\Delta_H(s) - 1}{\Delta_H(s) - 1} N_0^{\Delta_H(s)}}{\prod_{i=1}^{\Delta_H(s)} \lambda_i^{(s)}} \right]^{L_r} \tag{9}$$

where $\vec{\lambda}^{(s)}$ are the nonzero eigenvalues of the signal matrix $C_s = RC_c$, $\Delta_H(s)$ is the number of nonzero eigenvalues, C_c is the covariance matrix of fading coefficients and R is a matrix of the form:

$$R = \begin{pmatrix} R_{11} & R_{12} & \cdots & R_{1L_t} \\ R_{21} & R_{22} & \cdots & R_{2L_t} \\ \vdots & \vdots & \ddots & \vdots \\ R_{L_t 1} & R_{L_t 2} & \cdots & R_{L_t L_t} \end{pmatrix} \tag{10}$$

Each element of R is a diagonal submatrices of size N_c by N_c. The diagonal entries of the submatrix R_{ij} are given as

$$R_{ij}(k) = \int_{(k-1)T}^{kT} \delta_i^*(t, \vec{\alpha}, \vec{\beta}) \delta_j(t, \vec{\alpha}, \vec{\beta}) \quad \text{for} \quad k = 1, \cdots, N_c \tag{11}$$

where

$$\delta_i(t, \vec{\alpha}, \vec{\beta}) = x_i(t, \vec{\alpha}_i) - x_i(t, \vec{\beta}_i) \tag{12}$$

In order to minimize the error probability the rank of C_s should be maximized and so is its determinant.

As interleaver cannot be used after CPM otherwise the phase continuity of CPM signals will be disrupted, it is reasonable to assume that the fading coefficients remain constant during one error path. Herein we consider the special case where the fading channels are statistically independent and number of transmit antennas is $L_t = 2$. Thus the signal matrix can be reduced to

$$C_s = \begin{bmatrix} \int_0^T |\delta_1(t,\vec{\alpha}_1,\vec{\beta}_1)|^2 dt & \int_0^T \delta_1^*(t,\vec{\alpha}_1,\vec{\beta}_1)\delta_2(t,\vec{\alpha}_2,\vec{\beta}_2)dt \\ \vdots & \vdots \\ \int_{(N_c-1)T}^{N_cT} |\delta_1(t,\vec{\alpha}_1,\vec{\beta}_1)|^2 dt & \int_{(N_c-1)T}^{N_cT} \delta_1^*(t,\vec{\alpha}_1,\vec{\beta}_1)\delta_2(t,\vec{\alpha}_2,\vec{\beta}_2)dt \\ \int_0^T \delta_1(t,\vec{\alpha}_1,\vec{\beta}_1)\delta_2^*(t,\vec{\alpha}_2,\vec{\beta}_2)dt & \int_0^T |\delta_2(t,\vec{\alpha}_2,\vec{\beta}_2)|^2 dt \\ \vdots & \vdots \\ \int_{(N_c-1)T}^{N_cT} \delta_1(t,\vec{\alpha}_1,\vec{\beta}_1)\delta_2^*(t,\vec{\alpha}_2,\vec{\beta}_2)dt & \int_{(N_c-1)T}^{N_cT} |\delta_2(t,\vec{\alpha}_2,\vec{\beta}_2)|^2 dt \end{bmatrix} \quad (13)$$

3.2 Code Structure

First we consider the delay diversity scheme (denoted DD) as proposed by Wittneben [4]. This scheme transmits the same information from both antennas simultaneously but with a delay of one symbol interval. It can be easily shown that in DD C_s is guaranteed to have full rank, but it may not achieve maximum coding gain. As we can see from the (13), there are several methods to guarantee C_s full rank. One approach is to make

$$\int_{(k-1)T}^{kT} |\delta_1(t,\vec{\alpha}_1,\vec{\beta}_1)|^2 dt \int_{(k-1)T}^{kT} |\delta_2(t,\vec{\alpha}_2,\vec{\beta}_2)|^2 dt \neq \left| \int_{(k-1)T}^{kT} \delta_1^*(t,\vec{\alpha}_1,\vec{\beta}_1)\delta_2(t,\vec{\alpha}_2,\vec{\beta}_2)dt \right|^2 \quad (14)$$

for a given $k \in \{1, \cdots, N_c\}$. According to Scharwtz Inequality this inequality holds as long as

$$\delta_1(t,\vec{\alpha}_1,\vec{\beta}_1) \neq C\delta_2(t,\vec{\alpha}_2,\vec{\beta}_2). \quad (15)$$

Here the difference between linear modulation and nonlinear modulation comes into play. In linear modulation, for a given k,

$$\delta_1(t,\vec{\alpha}_1,\vec{\beta}_1) = \alpha_1(k) - \beta_1(k) \quad (16)$$

$$\delta_2(t,\vec{\alpha}_2,\vec{\beta}_2) = \alpha_2(k) - \beta_2(k) \quad (17)$$

which means that we always have

$$\delta_1(t,\vec{\alpha}_1,\vec{\beta}_1) = C\delta_2(t,\vec{\alpha}_2,\vec{\beta}_2) \quad (18)$$

Therefore, the above inequality does not hold at all. While in CPM system, as both $\delta_1(t,\vec{\alpha}_1,\vec{\beta}_1)$ and $\delta_2(t,\vec{\alpha}_2,\vec{\beta}_2)$ are functions of time, (15) can be easily satisfied. We can use different CPM schemes over different transmit antennas, which may result in higher complexity or more bandwidth occupation. Another method is to use different mapping rules (denoted DM) over different antennas to achieve full diversity. Following is just an example to demonstrate this idea. In order to show the performance we need to pick some specific CPM scheme. Here we consider 4-ary CPFSK with $h = 0.5$. Consequently, the coding gain provided by delay diversity is 4.

If we use the mapping rule as in table 1:

$D(k)$	$D_1(k)$	$D_2(k)$
0	-3	-1
1	-1	+3
3	+1	-3
2	+3	+1

Table 1: Mapping rule of the DM space-time code, $L_t = 2$, 4-ary

The resulting coding gain is 5.5154. Compared with delay diversity there is a 1.3951dB improvement. Fig. 2 shows the performance of DD and DM. (In all our simulations, each frame consists of 130 transmissions out of each transmit antenna and we choose the normalized Doppler spread to be $f_D T = 0.01$). The improvement of DM is more obvious when receive diversity is available, which can be explained through (9). When there is no receive diversity, although the coding gain for DM is large, the multiplicity for the coding gain is also large and they are of the same significance to the error performance; when the receive diversity is available, the coding gain plays a more important role than the multiplicity due to the exponential index L_r.

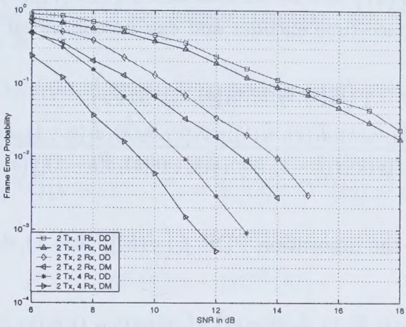


Figure 2: Performance comparison between DD and DM

4 Adaptive Soft-output Demodulation

In real application, CPM is very often combined with some outer coding scheme to achieve better performance. Actually the whole system shown in Fig. 1 can be treated as a serially concatenated system which we denote trellis-coded space-time CPM (TC-ST-CPM).

It is well known that a soft-output algorithm is necessary in a concatenated system for an iterative detection to be implemented. Here we consider K_d-lag soft-output demodulation of the space-time code in nonlinear modulation without perfect channel state information (CSI). The algorithm only assumes a knowledge of the statistical characteristics of the channel in forming the soft-output a posteriori probability (APP). While the CSI is not explicitly estimated, the algorithm implicitly estimates the CSI and uses this estimate in forming the desired APPs. The algorithms presented here are extensions of the work done in [5, 6, 7, 8]. Modifications of MLSD are proposed for CPM signals transmitted over frequency flat, Rayleigh fading channels in [5]. Seymour [6] gives the MAP algorithms for linearly modulated system in flat fading while Balasubramanian [7] extends these results to certain CPM signals. No either transmit diversity or receive diversity is considered in all these works. In [8] soft output receive diversity combining for CPM signals over space-time correlated Rayleigh fading channels is derived. Here we take both the transmit diversity and receive diversity into consideration.

The following notations are useful in deriving the algorithm. For any time series $\{f(0), f(1), \cdots, f(k)\}$, $\vec{f}(k) = [f(k) \ \cdots \ f(0)]^T$, boldface denotes signals over all receive antennas and $Y_j(k) = \{Y_j(t) : \quad kT \leq t < (k+1)T\}$. And as in general, capital letter indicates random variables while lower case indicates realization.

The optimum K_d-lag soft-output algorithm produces the metric $p\left(d(k-K_d)|\vec{\mathbf{y}}(k)\right)$ which can be written as

$$p\left(d(k-K_d)|\vec{\mathbf{y}}(k)\right) = \sum_{\{\theta_1(k),\theta_2(k),\vec{d}(k)\}\in\Gamma(d(k-K_d))} p\left(\theta_1(k),\theta_2(k),\vec{d}(k)|\vec{\mathbf{y}}(k)\right) \tag{19}$$

where $\{\theta(k),\theta_2(k),\vec{d}(k)\} \in \Omega_\Theta \times \Omega_D^k$ and $\Gamma(d(k-K_d)) = \{\Theta_1(k),\Theta_2(k),\vec{D}(k) : D(k-K_d) = d(k-K_d)\}$. The set $\Gamma(d(k-K_d))$ represents all possible phase trajectories such that $D(k-K_d) = d(k-K_d)$.

$p\left(\theta_1(k),\theta_2(k),\vec{d}(k)|\vec{\mathbf{y}}(k)\right)$ can be written into $p\left(\theta_1(k),\theta_2(k),\vec{d}_1(k),\vec{d}_2(k),|\vec{\mathbf{y}}(k)\right)$ by using the structure of space-time encoder and can be computed in an efficient recursive fashion as

$$\begin{aligned} &p\left(\theta_1(k),\theta_2(k),\vec{d}_1(k),\vec{d}_2(k)|\vec{\mathbf{y}}(k)\right) = \\ &\frac{f(\mathbf{y}(k)|\vec{\mathbf{y}}(k-1),\theta_1(k),\theta_2(k),\vec{d}_1(k),\vec{d}_2(k))p(\theta_1(k),\theta_2(k),\vec{d}_1(k-1),\vec{d}_2(k-1)|\vec{\mathbf{y}}(k-1))p(d(k))}{\sum_{\theta(k),\vec{d}(k)} f(\mathbf{y}(k)|\vec{\mathbf{y}}(k-1),\theta(k),\vec{d}(k))p(\theta(k),\vec{d}(k-1)|\vec{\mathbf{y}}(k-1))p(d(k))} \end{aligned} \tag{20}$$

The key part is to get the innovation PDF $f\left(\mathbf{y}(k)|\vec{\mathbf{y}}(k-1),\theta_1(k),\theta_2(k),\vec{d}_1(k),\vec{d}_2(k)\right)$ which can be written as

$$\int f\left(\mathbf{y}(k)|\vec{\mathbf{y}}(k-1),\theta_1(k),\theta_2(k),\vec{d}_1(k),\vec{d}_2(k),\mathbf{c}(k)\right) f\left(\mathbf{c}(k)|\vec{\mathbf{y}}(k-1),\theta_1(k),\theta_2(k),\vec{d}_1(k),\vec{d}_2(k)\right) d\mathbf{c}(k) \tag{21}$$

The first part can be defined using the Wiener measure [9] while the second part involves an extended Kalman filter [10] to estimate the mean and variance of $\mathbf{c}(k)$. Thus the innovation PDF can be computed through a multi-dimensional Gaussian integration. The whole algorithm is derived as follows:

$$f\left(\mathbf{y}(k)|\vec{\mathbf{y}}(k-1),\theta_1(k),\theta_2(k),\vec{d}_1(k),\vec{d}_2(k),\mathbf{c}(k)\right) = C_f \exp\left\{2\frac{E_s}{N_0}\Re\left[\mathbf{q}^H(k)\mathbf{c}(k)\right] - \frac{E_s}{N_0}\mathbf{c}^H(k)R(k)\mathbf{c}(k)\right\} \tag{22}$$

where

$$\begin{aligned}
\mathbf{q}(k) &= [q_{1,1}(k), q_{2,1}(k), \cdots, q_{1,L_r}(k), q_{2,L_r}(k)]^T && (23)\\
\mathbf{c}(k) &= [c_{1,1}(k), c_{2,1}(k), \cdots, c_{1,L_r}(k), c_{2,L_r}(k)]^T && (24)\\
q_{1,j}(k) &= \frac{1}{\sqrt{E_s T}}\int_{kT}^{(k+1)T} y_j(t)e^{-j\phi(t,d_1(k),\vec{s}_1(k))}dt && (25)\\
&= c_{1,j}(k-1) + \rho^*(k)c_{2,j}(k-1) + n_{1,j}(k-1) &&\\
q_{2,j}(k) &= \frac{1}{\sqrt{E_s T}}\int_{kT}^{(k+1)T} y_j(t)e^{-j\phi(t,d_2(k),\vec{s}_2(k))}dt && (26)\\
&= c_{2,j}(k-1) + \rho(k)c_{1,j}(k-1) + n_{2,j}(k-1) &&\\
R(k) &= diag\{A(k), \cdots, A(k)\} && (27)\\
A(k) &= \begin{bmatrix} 1 & \rho^*(k) \\ \rho(k) & 1 \end{bmatrix} && (28)\\
\rho(k) &= \frac{1}{T}\int_{kT}^{(k+1)T} e^{j\phi(t,d_1(k),\vec{s}_1(k))-j\phi(t,d_2(k),\vec{s}_2(k))}dt && (29)
\end{aligned}$$

Actually $q_{i,j}(k)$ is the matched filter output on the jth antenna for $[D_i(k),\vec{S}_i(k)] = [d_i(k),\vec{s}_i(k)]$. (Note for simplicity of notation we have dropped the dependence of $q_(i,j)(k)$

on $\theta_i(k)$, $\theta_j(k)$, $\vec{d_i}(k)$ and $\vec{d_j}(k)$). Notice there exists cross correlation between $q_{1,j}(k)$ and $q_{2,j}(k)$ ($\rho(k)$) even when the channels are independent due to the signal superposition from different transmitter antennas.

In order to compute the PDF $f\left(\mathbf{c}(k)|\vec{\mathbf{y}}(k-1), \theta_1(k), \theta_2(k), \vec{d_1}(k), \vec{d_2}(k)\right)$ we define the following random vectors.

$$\begin{aligned}\vec{Q}_j(k-1) &= [Q_{1,j}(k-1), \cdots, Q_{1,j}(0), Q_{2,j}(k-1), \cdots, Q_{2,1}(0)]^T &(30)\\ \vec{\mathbf{Q}}(k-1) &= \left[\vec{Q}_1^T(k-1), \cdots, \vec{Q}_{L_r}^T(k-1)\right]^T &(31)\end{aligned}$$

It is apparent that the past matched filter outputs are sufficient statistics for estimating $\mathbf{C}(k)$ and since $\mathbf{C}(k)$ and $\vec{\mathbf{Q}}(k-1)$ are jointly Gaussian, the PDF of $\mathbf{C}(k)$ conditioned on the past information is

$$f(\mathbf{c}(k)|\vec{\mathbf{y}}(k-1), \theta_1(k), \theta_2(k), \vec{d_1}(k), \vec{d_2}(k)) = \frac{1}{\pi^{2L_r}|\Sigma_e|} \exp\left\{-\left[\mathbf{c}(k) - \hat{\mathbf{c}}(k)\right]^H \Sigma_e^{-1} \left[\mathbf{c}(k) - \hat{\mathbf{c}}(k)\right]\right\} \quad (32)$$

where $\hat{\mathbf{c}}(k)$ and Σ_e are given as

$$\hat{\mathbf{c}}(k) = g^H \vec{\mathbf{Q}}(k-1) \qquad \Sigma_e = E\left[\mathbf{C}(k)\mathbf{C}^H(k)\right] - g^H P \quad (33)$$

in which

$$\begin{aligned}P &= RE\left[\vec{\mathbf{C}}(k-1)\mathbf{C}^H(k)\right] &(34)\\ \Sigma_c &= R^H E\left[\vec{\mathbf{C}}(k-1)\vec{\mathbf{C}}^H(k-1)\right] R &(35)\\ g &= \left(\Sigma_c + \frac{N_0}{E_s} I\right)^{-1} P &(36)\\ R &= diag\{A, \cdots, A\} &(37)\\ A &= \begin{bmatrix} I & B^H \\ B & I \end{bmatrix} &(38)\\ B &= diag\{\rho(k-1), \cdots, \rho(0)\} &(39)\end{aligned}$$

Actually (33) can be viewed as a minimum mean square error (MMSE) predictor which is a space-time processor. g serves as a $2L_r$ dimensional prediction filter of time length k, $\hat{\mathbf{c}}(k)$ are the predicted fading coefficients at time $t = kT$ and Σ_e is the prediction variance. Notice due to the signal cross correlation from different transmit antennas, both the filter and the variance are now functions of both channel self correlation and signal cross correlation and their evaluation cannot be carried out off-line, which is unlike the case when only receive diversity is available [8].

Closed form of the innovation PDF can be obtained as

$$f\left(\mathbf{y}(k)|\vec{\mathbf{y}}(k-1), \theta_1(k), \theta_2(k), \vec{d_1}(k), \vec{d_2}(k)\right) = N_f \times \quad (40)$$
$$\prod_{i=1}^{L_r} \exp\left(-\frac{E_s/N_0}{1 + E_s/N_0 \cdot 2\sigma_i^2}|z_i(k) - n_i(k)|^2 + \frac{E_s}{N_0}|z_i(k)|^2\right)$$

where

$$\begin{aligned}a\Sigma_e a^H &= U^H \begin{bmatrix} 2\sigma_1^2 & 0 & \cdots & 0 \\ 0 & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & 2\sigma_{L_r}^2 \end{bmatrix} U &(41)\\ \mathbf{n}(k) &= Ua\hat{\mathbf{c}}(k) \qquad \mathbf{z}(k) = Ua^{-1}\mathbf{q}(k) \qquad a^H a = R(k) &(42)\end{aligned}$$

The detector for the whole system is the same as that for standard serially concatenated codes [11] which is shown in Fig. 3. Based on the received signals from all antennas, $Y_1(t), \cdots, Y_{L_r}(t)$, and the a priori information, $p_e(D(k);i)$, the inner demodulator calculates the soft information $p_e(D(k);o)$. The outer decoder used this soft information to update the a priori information $p_e(D(k);i)$. This updated information is then used for the next iteration. After several iterations, the outer decoder calculates the soft information of the information bits, $p(I(\tilde{k});o)$ and makes decision.

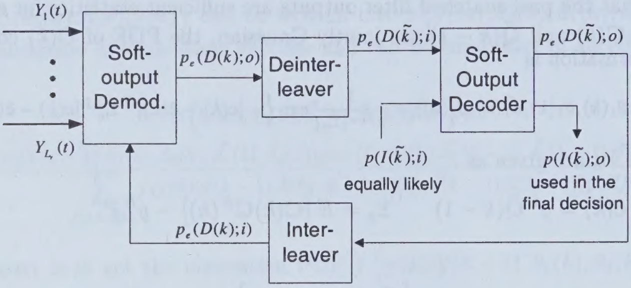

$Y_1(t)$
Soft-output Demod.
$Y_{L_r}(t)$
$p_e(D(k);o)$
Deinter-leaver
$p_e(D(\tilde{k});i)$
Soft-Output Decoder
$p_e(D(\tilde{k});o)$
$p(I(\tilde{k});i)$
equally likely
$p(I(\tilde{k});o)$
used in the final decision
$p_e(D(k);i)$
Inter-leaver

Figure 3: Iterative decoder structure

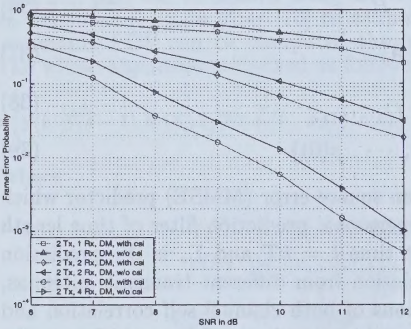


Figure 4: Simulation results of DM

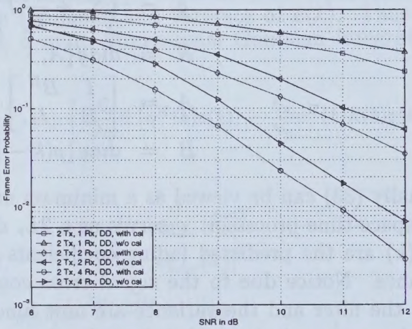


Figure 5: Simulation results of DD

Performance of the derived algorithm is shown in Fig. 4 and Fig. 5. (Note in order to fit the algorithm into practical applications we need to cut the channel memory into some finite length N_d). In these simulations, decision lag is chosen to be $K_d = 4$ and length of channel estimation filter is $N_d = 12$. There is no outer convolutional code considered here. The space-time codes used here are those given in section 3.2. These figures illustrate that the performance of the derived algorithm is close to that of perfect CSI with degradation being around 1dB when no receive diversity is available while the degradation being less when the receive diversity is available. Again the improvement of DM can be told through these two figures.

Performance of TC-ST-CPM is shown in Fig. 6. Here the outer channel encoder is a convolutional code with generator polynomial is $\bar{g}(5,7)$ and random interleaver is used.

The simulation results demonstrate several important points. First, by adding outer convolutional code, the whole system performance improves a lot. Second, the iterative detection provides significant improvement to the system performance even when CSI needs to be estimated. Third, although the CSI is generated within the inner space-time CPM demodulator, the exchange and update of soft information provides an implicit mechanism for the re-estimation of the unknown CSI. Therefore, the result is closer to that of perfect CSI compared to the case where there is no channel encoder.

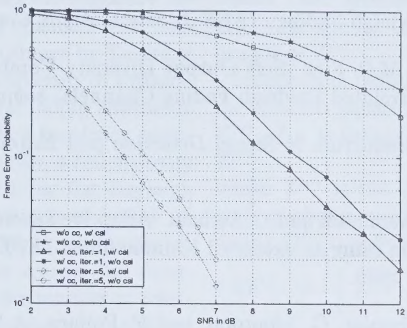


Figure 6: Simulation results of the TC-ST-CPM using DM

5 Conclusion

This paper gives the design criterion of space-time codes in CPM system. Its application to serially concatenated CPM system (TC-ST-CPM) is investigated. The adaptive soft-output demodulation without perfect channel state information is also presented. Simulation results show that both the space-time code and the derived adaptive demodulation algorithm give good performance.

References

[1] N.Seshadri V.Tarokh and A.R.Calderbank. Space-time codes for high data rate wireless communication: Performance criterion and code construction. *IEEE Trans. Inform. Theory*, Vol. 44:744–765, Mar 1998.

[2] T. Aulin J.B. Anderson and C.W.Sundberg. *Digital Phase Modulation.* Plenum, New York, 1986.

[3] M.P. Fitz, J. Grimm, and S. Siwamogsatham. A new view of performance analysis techniques in correlated Rayleigh fading. *IEEE Info. Theory*, Submitted 1999. Available at http://eewww.eng.ohio-state.edu/~fitz/Papers_folder/proberrj.ps.

[4] A.Wittneben. A new bandwidth efficient transmit antenna modulation diversity scheme for linear digital modulation. *Proc. IEEE'ICC*, pages 1630–1634, 1993.

[5] J.H.Lodge and M.L.Mother. Maximum likelihood sequence estimation of CPM signals transmitted over Rayleigh flat fading channels. *IEEE Trans. Commun.*, COM-38:784–794, June 1990.

[6] J.P.Seymour and M.P.Fitz. Near-Optimal Symbol-by-Symbol Detection Schemes for Flat Rayleigh Fading. *IEEE Trans. Commun.*, Vol. COM-41:1525–1533, February/March/April 1995.

[7] R. Balasubramanian and M.P. Fitz. Soft Output Detection of CPM Signals in Frequency Flat, Rayleigh Fading. *IEEE J. Select. Areas Commun.*, July 2000.

[8] Xiaoxia Zhang and M.P. Fitz. Soft Output Diversity Combining for CPM Signals over Space-time Correlated Rayleigh Fading Channels. *submitted to ICC*, 2001.

[9] H. V. Poor. *An Introduction to Signal Detection and Estimation.* Springer-Verlag, New York, 1994.

[10] H.Meyr A.Aghamohammadi and G.Ascheid. Adaptive synchronization and channel parameter estimation using an extended Kalman filter. *IEEE Trans. Commun.*, Vol. 37:1212–1219, Nov. 1989.

[11] S. Benedetto, D. Divsalar, G. Montorsi, and F. Pollara. A Soft-Input Soft-Output Maximum A Posteriori (MAP) Module to Decode Parallel and Serial Concatenated Codes. *TDA Progress Report*, Vol. 42-127, November 1996.

Space–Time Coding for Frequency Selective Fading Channel

Youjian Liu, Michael P. Fitz, and Oscar Y. Takeshita
Department of Electrical Engineering, The Ohio State University
`liuy@ee.eng.ohio-state.edu; {fitz.7, takeshita.3}@osu.edu`

Introduction: In this paper, we study the design of space–time codes for frequency selective quasi-static fading channels with non-negligible ISI. It has been shown by Tarokh *et al.* [1] that the achievable diversity of a space–time code designed for flat fading channel is not reduced if it is used in a frequency selective fading channel. However, a full diversity space-time code designed for flat Rayleigh fading channel may not be able to exploit all the available diversities. We propose a design methodology for codes to achieve full diversity $L_tL_rL_c$, where L_t is the number of transmit antennas, L_r is the number of receive antennas, and L_c is the number of independent ISI taps. First we introduce the concept of the virtual antenna. This idea permits a formalization of the performance criteria. We also propose to use algebraic design theory [2] to find STC that provide full diversity in an ISI channel.

Performance Criteria: Throughout the paper, we assume the number of receive antenna L_r be 1 because it does not affect performance criteria. Let N_c be frame size in terms of symbol time. A space-time code word $\mathbf{J}$ is an $N_c \times L_t$ matrix with elements drawn from a finite alphabet. The code symbol matrix $\mathbf{D}$ is defined as $\mathbf{D} = f(\mathbf{J})$, where $f(\bullet)$ is an element-wise constellation mapping from the finite alphabet to points of a constellation on the complex plane. The element at i^{th} column and j^{th} row of $\mathbf{D}$, $D_i(j)$, is transmitted from i^{th} antenna at time j.

In the scenario of quasi-static and time independent fading, if the fading coefficients are spatial independent with iid ISI taps, the performance criteria based on minimizing pairwise error probability can be proved to be maximizing diversity gain Δ_H and coding gain $\prod_{i=1}^{\Delta_H} \lambda_i$ [3]:

- Quasi-static fading: λ_i's are proportional to square of the nonzero singular values of the virtual code symbol difference matrix $\mathbf{Z}^{(v)}$. Δ_H is the number of nonzero singular values or rank of $\mathbf{Z}^{(v)}$;
- Independent fading: λ_i's are proportional to the nonzero square norm of row vectors of $\mathbf{Z}^{(v)}$. Δ_H is the number of nonzero rows of $\mathbf{Z}^{(v)}$;

where $\mathbf{Z}^{(v)}$ is the difference of two virtual code symbol matrices. A virtual code symbol matrix $\mathbf{D}^{(v)}$ is defined as

$$
\begin{aligned}
\mathbf{D}^{(v)} &= \Big[\ \vec{D}_1^{(0)} \ \ldots \ \vec{D}_{L_t}^{(0)} \ \ldots \\
&\qquad \vec{D}_1^{(L_c-1)} \ \ldots \ \vec{D}_{L_t}^{(L_c-1)}\ \Big], \\
\vec{D}_l^{(m)} &= \Big[\ \vec{0}_{1\times m} \ \ D_l(0) \ \ldots \\
&\qquad D_l(N_c-1) \ \ \vec{0}_{1\times(L_c-1-m)}\ \Big]^T.
\end{aligned}
$$

We observe that the virtual code symbol matrix $\mathbf{D}^{(v)}$ contains the original code symbol matrix $\mathbf{D}$ transmitted through the real antenna and the zero padded versions of $\mathbf{D}$'s transmitted through the virtual antennas. In the performance criterion of ISI channel, $\mathbf{D}^{(v)}$ plays the same role as $\mathbf{D}$ in non-ISI channel. For quasi-static fading channel, the maximum value of Δ_H is $L_t \times L_c$, which is true even if the independent ISI taps have unequal powers.

Code Design: We apply the powerful tool, Σ_o-rank theory [2] to design some full diversity example codes, whose coding gain is not necessarily maximized, to demonstrate the code design for fading ISI channel. By the concept of virtual antenna, the code design is reduced to the design of generator matrices of virtual code word. If all of the Σ_o linear combinations of the generator matrices have full Σ_o-rank, the code is guaranteed to achieve full diversity. For trellis codes without channel interleaver, the design is especially simple — only the generator polynomials need to be checked. For the case with channel interleaver, the code and the interleaver need to be jointly designed. Our approach is to select a trellis code first, then randomly construct the interleaver. If the code and interleaver pair does not satisfy the Σ_o-rank criterion, another interleaver is randomly chosen. The process is repeated until a full diversity code and interleaver pair is found.

For systems using trellis codes, with or without channel interleaver, the trellis codes need to satisfy the following constraints to achieve r-level diversity with 1 receive antenna. They are extensions of the corresponding constraints in [4].

Theorem 1: **(Without Channel Interleaver)** If a trellis code without channel interleaver is used in an independent L_c-tap ISI fading channel, to achieve transmission rate R bits/symbol and r level diversity with 1 receive antenna, the code must have at least $2^{R(r-L_c)}$ states and the constraint length is at least $r - L_c + 1$, where the constraint length is defined as minimum length of error paths of the original trellis code.

Theorem 2: **(With Channel Interleaver)** If a trellis code with channel interleaver is used in an independent L_c-tap ISI fading channel, to achieve transmission rate R bits/symbol and r level diversity with 1 receive antenna, the code must have at least $2^{R(r/L_c-1)}$ states and the constraint length is at least r/L_c.

Simulation Results: The performance of some example codes are evaluated by simulation for two cases. For systems with channel interleavers, iterative BCJR Soft Input Soft Output (SISO) equalizer and decoder are used. For systems without any channel interleaver, the decoder combine the ISI channel and trellis code to form a super trellis Viterbi or BCJR decoder.

The detailed example code structures and performance results are in [3]. We only show two of them in systems with 2 transmit antennas, 1 receive antenna in a 2-tap iid ISI quasi-static fading channel. Figure 1 compares the performance of two 16-state example codes with the outage probability. The solid line with circle is the performance of one code used in the system with a channel interleaver. The solid line with triangle is the performance of the other code used in the system without a channel interleaver. With channel interleaver, the performance improves about 1 dB. Both of them asymptotically have the same slopes as the outage probability. This implies that they achieve the desired 4-level diversity. Therefore, it demonstrates the effectiveness of the design procedure provided in this paper.

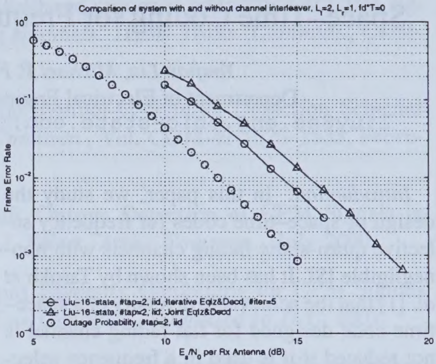


Fig. 1. Performance comparison of a code in the system with a channel interleaver and another code in the system without a channel interleaver. Both codes have 16 states.

Conclusions: We studied the space-time code design for fading ISI channel. By introducing the concept of virtual transmit antenna, the code design problem is translated to that of flat fading channel. Σ_o-rank theory is employed to design full diversity codes for systems with and without a channel interleaver. The simulation results show the example codes achieve the desired levels of diversity.

REFERENCES

[1] V. Tarokh, A. Naguib, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: performance criteria in the presence of channel estimation errors, mobility and multiple paths," *IEEE Trans. on Comm.*, vol. 47, no. 2, pp. 199–206, Feb. 1999.

[2] Y. Liu, M. P. Fitz, and O. Y. Takeshita, "A rank criterion for QAM space-time codes," *submitted to IEEE Trans. on Info. Theory*, Mar. 2000.

[3] Youjian Liu, *Space–time wireless communication*, Ph.D. thesis, The Ohio State University, Dec. 2000.

[4] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: performance criterion and code construction," *IEEE Trans. on Info. Th.*, vol. 44, no. 2, pp. 744–765, Mar. 1998.

Performance of Simple Space-Time Block Codes for More than Two Transmit Antennas

Balaji Raghothaman[1], Adrian Boariu[1], Olav Tirkkonen[2], and Ari Hottinen[2]
Nokia Research Center, [1]Irving, TX, USA; [2]Helsinki, Finland
email: balaji.raghothaman@nokia.com

Abstract

The performance of a non-orthogonal full rate space-time code that utilizes the Alamouti scheme is presented. In order to overcome the non-orthonormality, the decoding method involves iterative interference cancellation. Simulation results show that this scheme provides better diversity gain compared to the Alamouti scheme, for uncoded as well as coded data transmission in fading channels. Also, the proposed scheme can be easily adjusted for transmitting over three antennas.

1. Introduction

The concept of space-time codes (STC) for open-loop transmit diversity was introduced by Tarokh, *et al* [1]. The issue of decoding complexity for these codes was addressed by Alamouti [2]. From this work, the concept of space-time block coding emerged. The code suggested in [2] is a rate 1 code that achieves diversity 2 with a linear decoding scheme. The theory of space-time block codes was further developed by Tarokh, Jafarkhani and Calderbank [3]. They defined space-time block codes in terms of orthogonal code matrices. The properties of these matrices ensure full diversity equal to the number of transmit antennas, and a linear maximal likelihood detection. Theoretical results show that with one receive antenna, and perfect channel state information at the receiver, increasing the number of Tx antennas pays off at least up to 4. However, if a complex modulation alphabet is used, the cost of achieving 4-antenna transmit diversity with space-time block codes is reduced data rate. If a decrease in rate is not acceptable, one has to relax some other features of space-time block codes. First, one may sacrifice the uncoded diversity gain, and rely on coding to exploit the diversity provided by the additional antennas. Second, one may relax the requirement of orthogonality of the space-time block code [4]. In this paper the second approach is investigated.

2. General Notations and Alamouti Scheme

For the purpose of the presentation, it is assumed that the transmitter has L antennas and the receiver has one antenna. Let h_i, $i=1,2,\ldots,L$, denote the independent flat-fading channel impulse responses associated with the corresponding transmitting antenna. It is assumed that the channel coefficients are constant over L data symbols. Let $\mathbf{z} = [z_1, z_2, \ldots, z_L]^T$ be the L-tuple of consecutive data symbols, where T is the matrix transposition operator. These symbols will be mapped into the matrix $\mathbf{Z}$ for transmission. Since we are interested in a full rate STC (R=1), the matrix $\mathbf{Z}$ will be LxL square.

A. The Alamouti Scheme and its Properties

In the Alamouti scheme, L=2 and the pair of symbols $\mathbf{z}=[z_1, z_2]^T$ is mapped into the matrix, $\mathbf{Z}_{12} = \begin{bmatrix} z_1 & z_2 \\ -z_2^* & z_1^* \end{bmatrix}$, where the subscript of $\mathbf{Z}$ gives the data symbols indices involved in the mapping operation and * stands for the complex conjugation operator. The received signal is given by

$$\begin{bmatrix} r_1 \\ r_2 \end{bmatrix} = \mathbf{Z}_{12} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2 \end{bmatrix} \qquad (1)$$

In matrix $\mathbf{Z}_{12}$, the column index represents the antenna index, and the row index represents the time index. Equation (2) can be written equivalently as

$$\begin{bmatrix} r_1 \\ r_2^* \end{bmatrix} = \begin{bmatrix} h_1 & h_2 \\ h_2^* & -h_1^* \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} + \begin{bmatrix} n_1 \\ n_2^* \end{bmatrix} \Leftrightarrow \mathbf{r} = \mathbf{H}_{12}\,\mathbf{z} + \mathbf{n} \tag{2}$$

Note that there is a one-to-one correspondence between $\mathbf{Z}_{12}$ and $\mathbf{H}_{12}$. We will use the Alamouti STC as a sub-block for generating a STC for more than two antennas. Here, a property related to Alamouti's **H** matrix is given, which will play an important role in defining the STCs that will be proposed. Denoting by H the Hermitian operator, let's consider:

$$\mathbf{H}_{12}^H\,\mathbf{H}_{34} = \begin{bmatrix} x & y \\ -y^* & x^* \end{bmatrix},\ x = h_1^* h_3 + h_2 h_4^*,\ \text{and}\ y = h_1^* h_4 - h_2 h_3^*. \tag{3}$$

Note that

$$\mathbf{H}_{12}^H\,\mathbf{H}_{34} + \mathbf{H}_{34}^H\,\mathbf{H}_{12} = \left(x + x^*\right)\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},\ \mathbf{H}_{12}^H\,\mathbf{H}_{34} - \left(\mathbf{H}_{34}^H\,\mathbf{H}_{12}\right)^* = \left(y + y^*\right)\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}. \tag{4}$$

It is noticeable that the number of nonzero entries is halved in (4) compared to (3).

3. ABBA Schemes

Two simple constructions of STCs, for more than two transmit antennas, based on the Alamouti scheme, are described here. The first scheme, called ABBA, is given by

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{12} & \mathbf{H}_{34} \\ \mathbf{H}_{34} & \mathbf{H}_{12} \end{bmatrix};\ \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_{12} & \mathbf{Z}_{34} \\ \mathbf{Z}_{34} & \mathbf{Z}_{12} \end{bmatrix}, \tag{5}$$

while the second scheme, ABBA$_c$, has the following construction

$$\mathbf{H} = \begin{bmatrix} \mathbf{H}_{12} & \mathbf{H}_{34} \\ -\mathbf{H}_{34}^* & \mathbf{H}_{12}^* \end{bmatrix},\ \text{and}\ \mathbf{Z} = \begin{bmatrix} \mathbf{Z}_{12} & \mathbf{Z}_{34} \\ \mathbf{Z}_{34}^* & -\mathbf{Z}_{12}^* \end{bmatrix}. \tag{6}$$

The names ABBA and ABBA$_c$ come from the formation of the matrix **H** in (5) and (6) by the arrangement of the blocks A = $\mathbf{H}_{12}$ and B = $\mathbf{H}_{34}$. Note that in (5) and (6), we take advantage of the properties shown in (4).

A. ABBA decoder

For the sake of simplicity, only the decoder for ABBA scheme is presented. The decoder for the ABBA$_c$ scheme is similar. The received signal can be written in a form similar to (2)

$$\mathbf{r} = \alpha\,\mathbf{H}\,\mathbf{z} + \mathbf{n}, \tag{7}$$

where α stands for the ratio of the data channel amplitude to that of the pilot channel. Performing the matched filtering operation results in

$$\mathbf{y} = \mathbf{H}^H\,\mathbf{r} = \alpha\,\mathbf{H}^H\mathbf{H}\,\mathbf{z} + \mathbf{H}^H\mathbf{n} = \alpha\,\mathbf{M}\,\mathbf{z} + \mathbf{H}^H\mathbf{n}, \tag{8}$$

where $\mathbf{M} = \begin{bmatrix} a\,\mathbf{I}_2 & b\,\mathbf{I}_2 \\ b\,\mathbf{I}_2 & a\,\mathbf{I}_2 \end{bmatrix},\ \begin{matrix} a = \|h_1\|^2 + \|h_2\|^2 + \|h_3\|^2 + \|h_4\|^2 \\ b = x + x^*, x \text{ given in (4)} \end{matrix}.$ (9)

Note that **M** is a sparse matrix, which is a direct result of employing (4) in the code design. Also, z_1 interferes only with z_3, and z_2 interferes only with z_4. Left multiplying **y** by

$$\mathbf{M}^{(1)} = \begin{bmatrix} a\,\mathbf{I}_2 & -b\,\mathbf{I}_2 \\ -b\,\mathbf{I}_2 & a\,\mathbf{I}_2 \end{bmatrix} \text{ results in } \mathbf{M}^{(1)}\,\mathbf{y} = \alpha\left(a^2 - b^2\right)\mathbf{I}_4\,\mathbf{z} + \mathbf{M}^{(1)}\,\mathbf{H}^H\mathbf{n}. \tag{10}$$

Applying sign(·) function in (10) gives the data estimates of **z** (note $a \ge |b|$). Let $\hat{\mathbf{z}}_0$ be the initial estimate of the data vector **z.** These first data estimates can be further improved using a multistage interference cancellation scheme. The subsequent stage(s) can use **y** and the data estimates from previous stage to cancel the interference given by the b elements in the **M** matrix. The estimate after the i^{th} stage of interference cancellation is given by

$$\hat{\mathbf{z}}_i = dec\left(\mathbf{y} - \alpha\left(\mathbf{M} - a\,\mathbf{I}_4\right)\hat{\mathbf{z}}_{i-1}\right). \tag{11}$$

The decision function [*dec*(·)] can provide hard or soft outputs, the latter being desirable for further processing blocks, e.g. convolutional decoder.

Note that ABBA schemes can be tailored for three transmit antennas simply by omitting one of the antennas. For this case, in the **H** matrix the corresponding channel is set on zero, the decoding algorithm being the same.

4. Simulation Results

In this section, simulation results are presented, comparing ABBA to the Alamouti scheme. The bit error probability (BEP) curves for an uncoded system are shown in Figure 1, and those obtained for a coded CDMA2000 system are shown in Figure 2. The simulation parameters for the coded system are: chip rate = 1.2288 Mchips/sec., rate ¼ convolutional encoding, 9.6 kpbs data rate, no power control, and 74 Hz maximum Doppler shift. Figure 1 shows that at BEP = 10^{-2}, the ABBA scheme performs about 1 dB better than the Alamouti scheme. A similar gain is seen in Figure 2 for coded systems, at a BEP of 10^{-3}. The scheme denoted by ABBA+ABBA$_c$ in Figure 2 uses the two proposed schemes alternately for data block transmission. A slight performance gain can be achieved with this scheme compared to the situation in which only one scheme is used. The optimum 4 antenna diversity curves are also shown for reference. The non-orthogonality of the code is the major factor due to which the optimum performance is not achieved.

5. Conclusions

The performance of a non-orthogonal full rate space-time code for more than two antennas was presented. In order to overcome the non-orthonormality, a decoding method involving iterative interference cancellation was employed. Simulation results show that this scheme provides better diversity gain compared to the Alamouti scheme, for uncoded as well as coded data transmission in fading channels. Also, the proposed schemes can be easily adjusted for transmitting over three antennas.

References

[1] V. Tarokh, N. Seshadri, and A.R. Calderbank, "Space-time codes for high data rate wireless communication: Performance criterion and code construction," *IEEE Trans. Info. Theory*, Vol. 44, pp. 744-765, March 1998

[2] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE Journ. on Select Areas in Commun.*, vol. 16, no. 8, pp. 1451-1458, Oct. 1998

[3] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Trans. on Info. Theory*, vol. 45, pp. 1456-1467, July 1999

[4] O. Tirkkonen, A. Boariu, and A. Hottinen, "Minimal non-orthogonality rate 1 space-time block code for 3+ Tx antennas", *IEEE Int. Symp. on Spread Spectrum Tech. & Applications*, New Jersey, September 2000.

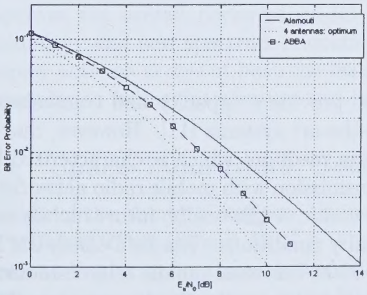


Figure 1. BEP for uncoded system

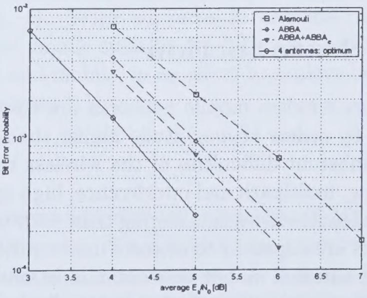


Figure 2. BEP for CDMA2000 system

An Interference and Fading Resilient Space-Time Coded CDMA Scheme

Hongbin Li
Department of Electrical and Computer Engineering
Stevens Institute of Technology
Hoboken, NJ 07030-5991, USA
hli@stevens-tech.edu

Jian Li
Department of Electrical and Computer Engineering
University of Florida
Gainesville, FL 32611-2044, USA
li@dsp.ufl.edu

Abstract

In this paper, we introduce a novel CDMA modulation scheme that integrates the strengths of space-time coding and spread spectrum technologies to achieve fading resiliency and multiuser interference suppression. Based on this new transmission scheme, several near-far resistant receivers are developed. Of particular interest are two differential receivers which do not require the channel state information (CSI) and are computationally very simple. The proposed modulation scheme, along with the novel receiver design, is able to work well in various fading environments ranging from stationary, quasi-stationary, to fast-fading. Additionally, both long and short spreading codes can be incorporated in this system. Simulation results are presented to illustrate the effectiveness of the proposed modulation scheme and receivers.

1 Introduction

Future wireless mobile networks are envisioned to provide a capacity and transmission rate by orders of magnitude higher than state-of-the-art systems [11]. However, due to the inherent difficulties of the wireless transmission medium, which is characterized by fading, multipath and interference, high-speed transmission with mobile radio networks is significantly more challenging than with wired networks. Additionally, future mobile networks are expected to operate under various mobility conditions. The IMT-2000/UMTS (international mobile telecommunications system 2000/universal mobile telecommunications system), currently being standardized and referred to as the third generation (3G) network, requires at least 144 Kb/s in a vehicular environment, 384 Kb/s in a pedestrian environment, and up to 2 Mb/s in an indoor environment [12]. Even more aggressive transmission rates with high mobility are being investigated for the fourth generation (4G) mobile network and beyond [2], [11]. With no doubt, rapidly time-varying fading will become a major concern in the design of future mobile networks.

The success of future high-speed mobile networks will decidedly depend on the ability to handle interference and fading. Making use of multiple transmit antennas, space-time coding provides an effective way to exploit spatial and temporal diversity, and is capable of drastically increase the channel capacity. A number of space-time coding and decoding schemes have been proposed [1], [15], [16]. However, most of these studies are based on the assumption that perfect channel state information (CSI) is available at the receiver, or at least can be accurately estimated by, for example, transmitting training symbols. Whereas training is a feasible approach to obtain the CSI when the channel is stationary or changes slowly, it will incur excessive overhead or even break down when the channel experiences fast fading. Most recently, differential space-time modulation schemes were introduced [6], [7], [14], which obviate channel estimation and, in the mean time, maintain the desired properties of space-time coding techniques. However, these schemes are interference-sensitive [8] and can not be used readily in CDMA systems, which are rich in multiuser interference due to all users overlapping in frequency and time.

In this paper, we propose a novel space-time coded CDMA transmission scheme which combines space-time coding and spread spectrum to suppress multiuser interference and combat fading. The proposed scheme does not require the CSI for detection, and can incorporate the CSI, if available, for enhanced performance. It is capable of functioning well in various mobile channels which may be stationary, quasi-stationary, or fast-fading (e.g., channel changing independently from symbol to symbol). Additionally, both short and long spreading codes can be incorporated in this scheme. Computer simulations are presented to demonstrate the effectiveness of the proposed modulation and detection techniques.

2 Signal Model

Consider a mobile communication system where multiples antennas are equipped at the base station and, possibly, at the mobile. Multiple antennas at the mobile are not required in the proposed scheme. However, if this becomes a possibility, then transmit and receive diversity can be implemented at both the base station and mobile, which significantly improves the overall performance (cf. Section 4). With advances in antenna design technology and new spectrum allocations, future mobile systems are fully anticipated to employ antenna arrays at both the base station and mobile, in an effort to achieve better immunity to fading/interference and higher capacity [11].

2.1 Space-Time Block Coding

In view of the above discussions, the mobile system under discussion consists of $M (M \geq 2)$ transmit antennas and $N (N \geq 1)$ receive antennas. We consider herein space-time block coding prior to transmission because of its ability to achieve full diversity with modest linear decoding complexity [1], [15]. To facilitate the presentation of the signal model of our proposed scheme, the basics of space-time block coding are briefly reviewed below.

Assume a baseband transmission employing a unit-energy signal constellation $\mathcal{C}$. At the transmitter, the data are encoded by M parallel encoders. Let $c_{ml} \in \mathcal{C}$ denote the constellation point selected by the encoder of transmit antenna $m = 1, \ldots, M$, at time $l = 1, \ldots, L$. The *code matrix* is defined as

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & \cdots & c_{1L} \\ c_{21} & c_{22} & \cdots & c_{2L} \\ \vdots & \vdots & \ddots & \vdots \\ c_{M1} & c_{M2} & \cdots & c_{ML} \end{bmatrix}, \tag{1}$$

where each column corresponds to the M symbols transmitted in parallel from the M transmit antennas. A space-time block code is a collection of $M \times L$ code matrices defined in the signal constellation $\mathcal{C}$.

Previous work on space-time coding assumes that the receiver has an accurate estimate of the channel state information (CSI) obtained by, e.g., training. Whereas training is a feasible approach to obtain the CSI when the channel is stationary or changes slowly, it will incur excessive overhead or even break down when the channel experiences fast fading.

Space-time block coding can be used for interference suppression under certain conditions. In particular, assuming that the system is equipped with *more receive antennas than the total number of users*, zero-forcing (ZF) and minimum mean squared error (MMSE) type of techniques can be implemented for interference suppression [10]. ZF requires the knowledge the CSI of all users, including the interfering users, while for the MMSE solution, only the CSI of the desired user is needed. In either case, interference suppression based on space-time block coding becomes much more difficult when time-varying fading channels are involved.

In order to deal with rapidly time-varying Rayleigh environments, where the CSI are unknown and cannot be reliably estimated at the receiver, Marzetta and Hochwald [5], [9] suggested to choose code matrices with equal-energy, orthogonal rows: $\mathbf{C}\mathbf{C}^H = L\mathbf{I}_M$, where $(\cdot)^H$ denotes the Hermitian transpose and $\mathbf{I}_M$ is the $M \times M$ identity matrix, in order to approach the channel capacity. This leads to the so-called *unitary space-time modulation* signaling [5].

Differential space-time modulation (DSTM) without requiring the CSI at the receiver is being under intensive research today. A number of different schemes have been proposed [6], [7], [14]. Among them, the approach by Hughes makes use of *unitary group codes* and can be applied to any number of antennas and to any constellation [7]. Specifically, by assuming $M = L$, unitary group codes are used to yield a simple DSTM scheme in [7]. For example, the code

$$\mathcal{G} = \left\{ \pm \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \pm \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix} \right\}, \quad \mathbf{D} = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \tag{2}$$

is a unitary group code of dimension two over the binary phase-shift keying (BPSK) constellation [7]. Each information bit pair in $\{00, 01, 10, 11\}$ corresponds to a space-time code in $\mathcal{G}$. To send the message $\mathbf{G}_t \in \mathcal{G}$ at time t, the transmitter sends $\mathbf{C}_t \in \mathcal{C}^{M \times L}$, where

$$\mathbf{C}_t = \mathbf{C}_{t-1}\mathbf{G}_t, \tag{3}$$

and the transmitter begins with $\mathbf{C}_0 = \mathbf{D}$. Note the analogy between (3) and the standard scalar-based differential phase-shift keying (DPSK) [13]. For the above BPSK example,

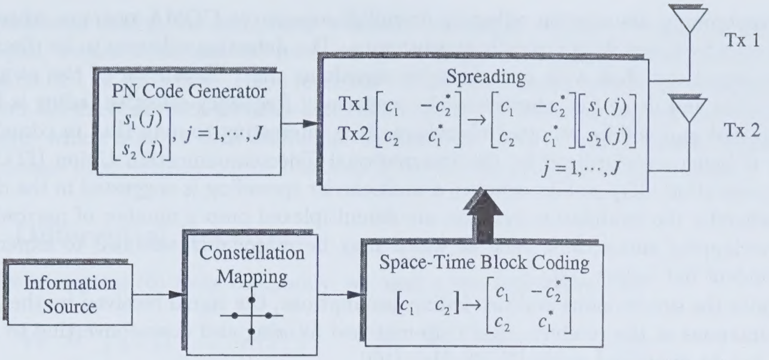

Tx 1
Tx 2
PN Code Generator
$\begin{bmatrix} s_1(j) \\ s_2(j) \end{bmatrix}, j=1,\cdots,J$
Spreading
Tx1 $\begin{bmatrix} c_1 & -c_2^* \\ c_2 & c_1^* \end{bmatrix} \rightarrow \begin{bmatrix} c_1 & -c_2^* \\ c_2 & c_1^* \end{bmatrix}\begin{bmatrix} s_1(j) \\ s_2(j) \end{bmatrix}$
Tx2
$j=1,\cdots,J$
Information Source
Constellation Mapping
Space-Time Block Coding
$[c_1 \ c_2] \rightarrow \begin{bmatrix} c_1 & -c_2^* \\ c_2 & c_1^* \end{bmatrix}$

Figure 1: Baseband space-time coding and spreading modulation.

two transmit antennas are used to transmit each of the two elements of one column of $\mathbf{C}_t$ at a time. The unitary group code matrices are orthogonal [7], i.e.,

$$\mathbf{C}_t\mathbf{C}_t^H = \mathbf{C}_t^H\mathbf{C}_t = M\mathbf{I}_M, \ \forall t, \tag{4}$$

where we have assumed $M = L$.

The DSTM scheme in [7] works very well in time-varying fading channels, provided that the channel coefficient do not change significantly over two consecutive symbol intervals, i.e., the time corresponding to transmitting two successive code matrices. However, the DSTM scheme was found to be sensitive to interference unknown to the receiver [8].

2.2 A Novel Spreading Modulation Scheme

Space-time coding and modulation as discussed above are sensitive to interference and, therefore, cannot be readily applied to CDMA systems. In the sequel, we introduce a new modulation scheme that integrates the merits of space-time coding and spread spectrum technologies for interference suppression. Figure 1 illustrates the baseband representation for this scheme with two transmit antennas. The idea is to *vector-spread and combine* each column of a space-time block code matrix. Specifically, each column of $\mathbf{C}_t$ is spread temporally with a unique spreading code of unit energy and the coded columns of $\mathbf{C}_t$ are transmitted *at the same time*. Let the output of the spreading modulator corresponding to the tth block and jth chip be $\mathbf{x}_t(j)$. Then it can be written as

$$\mathbf{x}_t(j) = \mathbf{C}_t\mathbf{s}(j), \ j = 1, \ldots, J, \tag{5}$$

where $\mathbf{s}(j)$, $j = 1, \ldots, J$, are $L \times 1$ vectors, corresponding to L spreading codes of length J. Here, we do not restrict $\mathbf{s}(j)$ to be periodic. Hence, *both short and long spreading codes can be used in this modulation scheme*. The spread signal $\mathbf{x}_t(j)$ is next pulse-shaping filtered and carrier modulated before being fed into the M transmit antennas for transmission.

2.3 Channel Model

We consider a synchronous K-user system operating in a Rayleigh flat-fading channel. The fading coefficients may either change slowly or vary rapidly from symbol to symbol.

The synchronous assumption reflects a downlink scenario in CDMA systems, where the signal seen by a mobile is typically synchronous. The detection schemes to be discussed in the sequel can deal with asynchronous signals as well. Extension of the proposed modulation and detection schemes to the scenario of frequency-selective fading is being investigated and will be reported elsewhere. It is interesting to note that in cdma2000, which is being standardized by the International Telecommunication Union (ITU) for third-generation (3G) mobile systems, a multicarrier spreading is suggested in the downlink, whereby the modulation symbols are demultiplexed onto a number of narrowband non-overlapping subcarriers, each of which may be reasonably assumed to experience independent flat fading [12].

Under the synchronous and flat-fading assumptions, the signal received by the N receive antennas at the receiver, after chip-matched filtering and downconverting to baseband, can be expressed as [5]–[7], [9], [14]–[16]

$$\mathbf{y}_t(j) = \sum_{k=1}^{K} \sqrt{\rho_{k,t}}\mathbf{H}_{k,t}\mathbf{B}_{k,t}\mathbf{s}_k(j) + \mathbf{n}_t(j), \ j = 1, \ldots, J, \tag{6}$$

where $\rho_{k,t}$ denotes the received signal power due to user k at time t per transmit per receive antenna, $\mathbf{H}_{k,t} \in \mathbb{C}^{N \times M}$ is the channel matrix for user k at time t, $\mathbf{B}_{k,t}$ is the space-time code matrix for user k at time t (to be specified shortly), $\mathbf{s}_k(j)$ denote the spreading codes for user k, and $\mathbf{n}_t(j) \in \mathbb{C}^{N \times 1}$ is the channel noise at time t. Note that both t and j are time indexes, with the former denoting the tth block code and the latter as the jth chip of the tth block code. We assume that the elements in $\mathbf{H}_{k,t}$ and $\mathbf{n}_t(j)$ are independent, identically-distributed, complex Gaussian random variables with zero mean and unit variance [5]–[7], [9], [14]–[16]. The signal-to-noise ratio (SNR) per receive antenna for user k and at time t is defined as $M\rho_{k,t}$. Note that the channel matrices $\{\mathbf{H}_{k,t}\}$ are different from one another for the uplink; however, for the downlink, $\mathbf{H}_{1,t} = \ldots = \mathbf{H}_{K,t}$.

If no channel state information (CSI) is available at the receiver (due to, e.g., fast fading), $\mathbf{B}_{k,t} \in \mathbb{C}^{M \times 2L}$ consists of two differential space-time code matrices:

$$\mathbf{B}_{k,t} \triangleq \left[\ \mathbf{C}_{k,t-1} \quad \mathbf{C}_{k,t} \ \right], \tag{7}$$

where $\mathbf{C}_{k,t-1}$ and $\mathbf{C}_{k,t}$ are $M \times L$ matrices to be transmitted at time $t-1$ and t and are differentially coded [cf. (3)]. If the CSI is available (for time-invariant or slow fading channels), then no differential coding is necessary and $\mathbf{B}_{k,t} \in \mathbb{C}^{M \times L}$ can be any space-time block codes in, e.g., [15].

The problem of interest is to determine the coding matrices $\mathbf{B}_{k,t}, k = 1, \ldots, K$, or, equivalently, the transmitted symbols for K users from the received data $\{\mathbf{y}_t(j)\}_{j=1}^{J}$ at time t.

3 Receiver Design

Given the data model in (6), standard detection schemes such as the optimum nonlinear maximum likelihood (ML) receiver, linear zero-forcing (ZF) receiver, and minimum mean squared error (MMSE) receiver are easily obtained. However, the ML receiver will have an exponential complexity, whereas the ZF and MMSE receivers will have difficulty in dealing with (fast) time-varying channel fading, similar to their counterparts that do not use space-time coding. As such, these standard techniques will not be pursued here. In the sequel, we first introduce two receivers which employ differential unitary group codes [7] and no CSI for detection. These receivers will be customarily referred to as *differential receivers*. In the case when accurate estimation of the CSI becomes feasible, we introduce a receiver which exploits such additional information. As one would expect, significant performance improvement can be obtained if the CSI is available to the receiver.

3.1 Differential Receivers

In order to express (6) more compactly, we need a few definitions. Let

$$\begin{array}{llll} \mathbf{Y}_t \triangleq \left[\ \mathbf{y}_t(1)\ \ldots\ \mathbf{y}_t(J)\ \right], & \mathbf{N}_t \triangleq \left[\ \mathbf{n}_t(1)\ \ldots\ \mathbf{n}_t(J)\ \right], \\ \mathbf{S}_k \triangleq \left[\ \mathbf{s}_k(1)\ \ldots\ \mathbf{s}_k(J)\ \right], & \mathbf{S} \triangleq \left[\ \mathbf{S}_1^T\ \ldots\ \mathbf{S}_K^T\ \right]^T, \\ \mathbf{A}_t \triangleq \left[\ \mathbf{A}_{1,t}\ \ldots\ \mathbf{A}_{K,t}\ \right], & \mathbf{A}_{k,t} \triangleq \sqrt{\rho_{k,t}}\mathbf{H}_{k,t}\mathbf{B}_{k,t},\ k = 1, \ldots, K, \end{array} \tag{8}$$

where $(\cdot)^T$ denotes the transpose. Then (6) can be expressed as

$$\mathbf{Y}_t = \mathbf{A}_t\mathbf{S} + \mathbf{N}_t. \tag{9}$$

Note that all information about the transmitted symbols $\{\mathbf{B}_{k,t}\}_{k=1}^K$ is contained in $\mathbf{A}_t$. If an accurate estimate of $\mathbf{A}_t$ can be obtained, then hopefully we may use it for reliable detection of the symbols. This is exactly the approach we will be pursuing.

The *unstructured* maximum likelihood (ML) estimate of $\mathbf{A}_t$, conditioned on $\{\mathbf{H}_{k,t}\}_{k=1}^K$ and $\{\mathbf{B}_{k,t}\}_{k=1}^K$, is given by

$$\hat{\mathbf{A}}_t = \mathbf{Y}_t\mathbf{S}^H \left(\mathbf{S}\mathbf{S}^H\right)^{-1}, \tag{10}$$

where we have assumed that $\mathbf{S}\mathbf{S}^H$ has full rank, and by "unstructured", we mean the structure of $\mathbf{A}_t$ as specified in (8) is not enforced in the estimation. Due to the block structure of $\mathbf{A}_t$, the ML estimate, $\hat{\mathbf{A}}_{k,t}$, of $\mathbf{A}_{k,t}$ is easily obtained from $\hat{\mathbf{A}}_t$. Note again that the information about the transmitted symbol for user k is solely contained in $\mathbf{A}_{k,t}$, not in the other blocks of $\mathbf{A}_t$.

In order to use $\hat{\mathbf{A}}_{k,t}$ for detecting user k, we need to find out its statistics. It can be shown, after some calculations, that $\mathrm{vec}(\hat{\mathbf{A}}_{k,t})$, where $\mathrm{vec}(\cdot)$ denotes the operation of stacking the columns of a matrix on top of one another [4], is complex Gaussian with zero mean and covariance matrix

$$\Sigma_k = M\hat{\rho}_{k,t}\left(\left[\begin{array}{cc} \mathbf{I}_M & \mathbf{G}_{k,t} \\ \mathbf{G}_{k,t}^H & \mathbf{I}_M \end{array}\right]\right)^* \otimes \mathbf{I}_N + \mathbf{\Phi}_k, \tag{11}$$

where $(\cdot)^*$ denotes complex conjugate, $\mathbf{\Phi}_k \in \mathbb{R}^{2LN\times 2LN}$ is the kth diagonal block of $\mathbf{\Phi}$:

$$\mathbf{\Phi} \triangleq \left(\mathbf{S}\mathbf{S}^H\right)^{-1} \otimes \mathbf{I}_N, \tag{12}$$

with $\otimes$ denoting the matrix Kronecker product [4], and $\hat{\rho}_{k,t}$ is any consistent estimate of $\rho_{k,t}$, one given by

$$\hat{\rho}_{k,t} = \frac{1}{2M^2N}\,\mathrm{tr}(\hat{\mathbf{A}}_{k,t}\hat{\mathbf{A}}_{k,t}^H). \tag{13}$$

Given the distribution, receivers are straightforward to obtain.

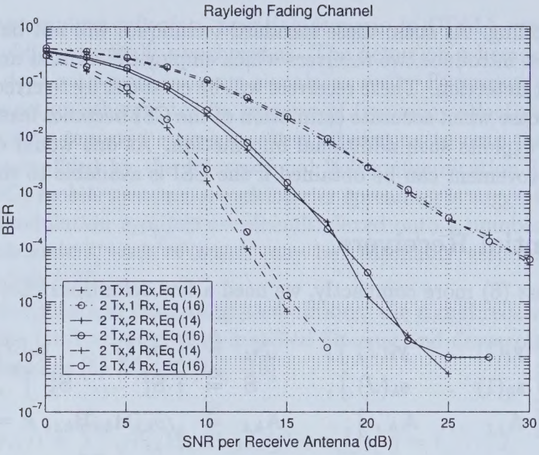


Figure 2: BER vs. SNR for the two differential receivers, described by (14) and (16), in a fast-fading Rayleigh channel.

3.1.1 Differential Receiver # 1

Applying the maximum likelihood (ML) methodology, we maximize the distribution of $\mathrm{vec}(\hat{\mathbf{A}}_{k,t})$, as derived in the above, with respect to the unknown unitary group code matrix $\mathbf{G}_{k,t}$, which gives the following receiver:

$$\hat{\mathbf{G}}_{k,t} = \arg\min_{\mathbf{G}_{k,t}\in\mathcal{G}} \left\{ \ln|\Sigma_k| + \mathrm{vec}^H(\hat{\mathbf{A}}_k)\Sigma_k^{-1}\mathrm{vec}(\hat{\mathbf{A}}_k) \right\},\ k = 1,\ldots,K. \tag{14}$$

3.1.2 Differential Receiver # 2

An idea that simplifies the receiver (14) quite a bit is to use the fact that the spreading sequences are typically chosen to have small cross correlation, which implies that $\mathbf{\Phi}_k$ is approximately equal to $\mathbf{I}_{2LN}$. We can show that, within a second-order approximation, $\mathbf{\Phi}_k$ can be replaced by $\mathbf{I}_{2LN}$ in (11) without affecting the asymptotic (meaning for a not too small spreading gain J) performance. Doing so yields a significantly simplified receiver. Specifically, let $\hat{\mathbf{A}}_{k,t}^{(1)} \in \mathbb{C}^{N\times L}$ and $\hat{\mathbf{A}}_{k,t}^{(2)} \in \mathbb{C}^{N\times L}$ be defined as

$$\left[\begin{array}{cc} \hat{\mathbf{A}}_{k,t}^{(1)} & \hat{\mathbf{A}}_{k,t}^{(2)} \end{array}\right] \triangleq \hat{\mathbf{A}}_{k,t}. \tag{15}$$

The new receiver is described by

$$\hat{\mathbf{G}}_{k,t} = \arg\min_{\mathbf{G}_{k,t}\in\mathcal{G}} \left\{ \mathrm{Re}\left[\mathrm{tr}\left(\mathbf{G}_{k,t} (\hat{\mathbf{A}}_{k,t}^{(2)})^H \hat{\mathbf{A}}_{k,t}^{(1)} \right) \right] \right\},\ k = 1,\ldots,K, \tag{16}$$

where Re[·] takes the real-valued part of the argument and tr(·) denotes the matrix trace function. This receiver closely resembles the differential space-time modulation (DSTM) receiver in [7]. The only nominally added complexity of the current receiver comes from the calculation of $\hat{\mathbf{A}}_{k,t}$.

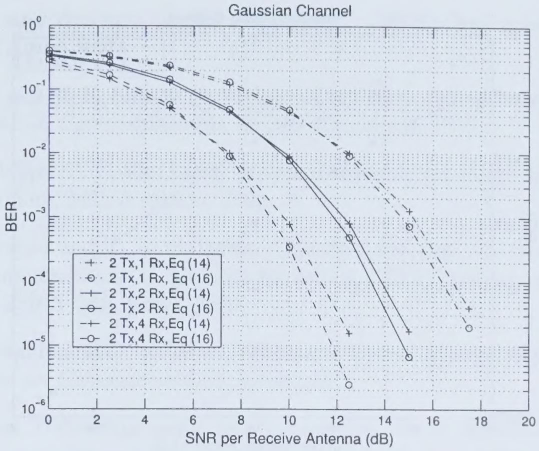


Figure 3: BER vs. SNR for the two differential receivers, described by (14) and (16), in a stationary Gaussian channel.

3.2 Another Receiver Assuming CSI

Let $\mathbf{Y}_t$, $\mathbf{A}_t$, $\{\mathbf{A}_{k,t}\}_{k=1}^{K}$ and $\mathbf{S}$ be similarly defined as in Section 3.1, except that $\{\mathbf{B}_{k,t} \in \mathbb{C}^{M\times L}\}_{k=1}^{K}$ and no differential coding is performed. Likewise, only L (instead of $2L$ in the previous case) spreading codes are needed for each user, reducing the dimension of $\mathbf{S}$ to $LK \times J$. The unstructured ML estimate $\hat{\mathbf{A}}_t$ of $\mathbf{A}_t$, conditioned on $\{\mathbf{H}_{k,t}\}_{k=1}^{K}$ and $\{\mathbf{B}_{k,t}\}_{k=1}^{K}$, has the same form as (10), except for the difference in the dimensions of the involved matrices. The ML estimate $\hat{\mathbf{A}}_{k,t}$ of $\mathbf{A}_{k,t}$ can again be obtained from $\hat{\mathbf{A}}_t$. In the current case, the CSI is known and no longer modeled as random. The conditional distribution of $\mathrm{vec}(\hat{\mathbf{A}}_{k,t})$ is complex Gaussian with mean $\mathrm{vec}(\mathbf{A}_{k,t})$ and covariance matrix $\mathbf{\Phi}_k$, where $\mathbf{\Phi}_k \in \mathbb{R}^{LN\times LN}$ is the kth diagonal block of $\mathbf{\Phi} \triangleq \left(\mathbf{S}\mathbf{S}^H\right)^{-1} \otimes \mathbf{I}_N$. Maximizing the distribution of $\mathrm{vec}(\hat{\mathbf{A}}_{k,t})$ with respect to the coding matrix $\mathbf{B}_{k,t}$ yields the following receiver:

$$\hat{\mathbf{B}}_{k,t} = \arg\min_{\mathbf{B}_{k,t}\in\mathcal{C}} \left\{ \left[\mathrm{vec}(\hat{\mathbf{A}}_{k,t}) - \mathrm{vec}(\mathbf{A}_{k,t})\right]^H \mathbf{\Phi}_k^{-1} \left[\mathrm{vec}(\hat{\mathbf{A}}_{k,t}) - \mathrm{vec}(\mathbf{A}_{k,t})\right] \right\}, \quad k = 1, \ldots, K. \tag{17}$$

4 Simulations

We consider a $K = 10$ synchronous CDMA system which is equipped with $M = 2$ transmit antennas and $N = 1, 2$ or 4 receive antennas. A BPSK constellation is employed for modulation. A near-far environment is simulated, with all nine interfering users transmitting at a power level 10 dB larger than that of the desired user, referred to as user 1. Gold codes with $J = 63$ are used for spreading. Both a stationary Gaussian and a fast-fading Rayleigh channels are simulated. In the latter case, the fading coefficients vary independently from one code block period to another. Figure 2 illustrates the the bit-error-rate (BER) of the two differential receivers described by (14) and (16), respectively,

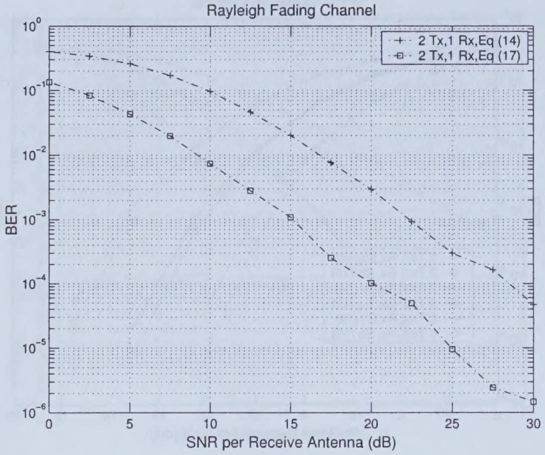


Figure 4: BER vs. SNR for the differential receiver, described by (14), and the receiver with perfect CSI, described by (17), in a fast-fading Rayleigh channel.

as a function of the SNR for user 1, which is defined as SNR $= M\rho_{1,t}$, when Rayleigh fading occurs. Figure 3 is the corresponding results with the Gaussian channel. It is noted that the two differential receivers perform similarly for both the Gaussian and Rayleigh fading channels. Hence, *the computationally simpler receiver (16) should be preferred in practice*. Observe that increasing the number of receive antennas from $N = 1$ to $N = 2$ or from $N = 2$ to $N = 4$ leads to a performance improvement of approximately 3 dB (in SNR) for the Gaussian channel [cf. Figure 3], which is not unexpected since in either case, the total received signal power is doubled. Note, however, that the BER improvement obtained by increasing the number of receive antennas is far more striking for the Rayleigh fading channel. For example, an improvement of approximately 10 dB is observed at BER $= 10^{-4}$ when N increases from 1 to 2 [cf. Figure 2]! This suggests that *whenever possible, receive antenna diversity should be pursued in a Rayleigh fading environment*.

Figure 4 compares the performance of the differential receiver (14) and the receiver with perfect CSI, described by (17), when only $N = 1$ receive antenna is used. We observe a considerable performance loss is observed when no CSI is available. The performance loss of the differential receiver may be reduced by including more symbols for joint detection, similar to use multiple symbols for detecting conventional differential phase-shift keying (DPSK) signals [3]. This extension is being investigated and will be reported in the near future.

References

[1] S. M. Alamouti. A simple transmit diversity techniques for wireless communications. *IEEE Journal on Selected Areas in Communications*, 16(8):1451–1458, October 1998.

[2] J. Chuang and N. Sollengerger. Beyond 3G: Wideband wireless data access based

on OFDM and dynamic packet assignment. *IEEE Communications Magazine*, 38(7):78–87, July 2000.

[3] D. Divsalar and M. K. Simon. Multiple symbol differential detection of DPSK. *IEEE Transactions on Communications*, 38:398–400, March 1990.

[4] G. H. Golub and C. F. Van Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore, MD, 3 edition, 1996.

[5] B. M. Hochwald and T. L. Marzetta. Unitary space-time modulation for multiple-antenna communications in Rayleigh flat fading. *IEEE Transactions on Information Theory*, 46(2):543–564, March 2000.

[6] B. M. Hochwald and W. Sweldens. Differential unitary space-time modulation. *IEEE Transactions on Communications*, submitted.

[7] B. L. Hughes. Differential space-time modulation. *Proc. 1999 IEEE Wireless Communications and Networking Conference (WCNC9́9),* New Orleans, LA, September 22-29 1999.

[8] J. Li, J. Liu, and H. Li. Differential space-code modulation for interference suppression. *Proceedings of 34th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, October 29 – November 1, 2000.

[9] T. L. Marzetta and B. M. Hochwald. Capacity of a mobile multiple-antenna communication link in Rayleigh flat-fading. *IEEE Transactions on Information Theory*, 45(1):139–157, January 1999.

[10] A. F. Naguib, N. Seshadri, and A. R. Calderbank. Increasing data rate over wireless channels. *IEEE Signal Processing Magazine*, 17(3):76–92, May 2000.

[11] NSF. Tetherless T3 and beyond. Interim Report, November 1998.

[12] R. Prasad and T. Ojanperä. An overview of CDMA evolution toward wideband CDMA. *IEEE Communications Surveys*, 1(1):2–29, Fourth Quarter 1998.

[13] J. G. Proakis. *Digital Communications.* McGraw-Hill, 3 edition, 1995.

[14] V. Tarokh and H. Jafarkhani. A differential detection scheme for transmit diversity. *IEEE Journal on Selected Areas in Communications*, 2000, submitted.

[15] V. Tarokh, H. Jafarkhani, and A. R. Calderbank. Space-time block codes from orthogonal designs. *IEEE Transactions on Information Theory*, 45(5):1456–1467, July 1999.

[16] V. Tarokh, N. Seshadri, and A. R. Calderbank. Space-time codes for high data rate wireless communications: Performance criterion and code construction. *IEEE Transactions on Information Theory*, 44(2):744–765, March 1998.

Designs in Additive Codes over $GF(4)$

Jon-Lark Kim and Vera Pless
Department of Mathematics, Statistics, and Computer Science,
322 SEO(M/C 249),
University of Illinois–Chicago,
851 S. Morgan, Chicago, IL 60607-7045, USA
e-mail: jlkim@math.uic.edu, pless@math.uic.edu

Abstract

The purpose of this paper is to study designs in additive codes over $GF(4)$. There are two types of designs. One is a classical t-design with repeated blocks. For these we have an analogue of the Assmus-Mattson Theorem for additive codes over $GF(4)$. The other is a generalized t-design first introduced by Delsarte [4]. As an example, we consider the dodecacode, which is the unique additive even self-dual $(12, 2^{12}, 6)$ code. We show that there exists a 5-$(12, 6, 3)$ design with multiplicity 1 or 3 in the dodecacode. We also derive a new simple 3-design, i.e. a 3-$(11, 5, 4)$ design from the shortened dodecacode.

Keywords Additive code, Assmus-Mattson Theorem, generalized t-design.

1 Introduction

Recently additive codes over $GF(4)$ have gained in popularity because of their connection to quantum codes [3]. Several papers [2, 5, 6, 7] were devoted to classifing additive self-dual codes over $GF(4)$. It was shown in [7] that certain vectors in some additive self-dual codes over $GF(4)$ hold a generalized t-design. We study further a generalized t-design in additive codes over $GF(4)$ as well as classical t-designs with repeated blocks. We give some basic definitions.

An *additive code* $\mathcal{C}$ *over* $GF(4)$ *of length* n is an additive subgroup of $GF(4)^n$. As $\mathcal{C}$ is a free $GF(2)$-module, it has size 2^k for some $0 \leq k \leq 2n$. We call $\mathcal{C}$ an $(n, 2^k)$ code. It has a basis, as a $GF(2)$-module, consisting of k basis vectors; a *generator matrix* of $\mathcal{C}$ will be a $k \times n$ matrix with entries in $GF(4)$ whose rows are a basis of $\mathcal{C}$. Interest in additive codes over $GF(4)$ has arisen because of their correspondence to quantum codes as described in [3]. There is a natural inner product arising from the trace map. If we let $GF(4) = \{0, 1, \omega, \overline{\omega}\}$ where $\overline{\omega} = \omega^2 = 1 + \omega$, the *trace* map $\text{Tr} : GF(4) \to GF(2)$ is given by

$$\text{Tr}(x) = x + x^2.$$

In particular $\text{Tr}(0) = \text{Tr}(1) = 0$ and $\text{Tr}(\omega) = \text{Tr}(\overline{\omega}) = 1$. The *conjugate* of $x \in GF(4)$, denoted $\overline{x}$, is the image of x under the Frobenius automorphism; in other words, $\overline{0} = 0$

$\overline{1} = 1$, and $\overline{\overline{\omega}} = \omega$. We now define the *trace inner product* of two vectors $\mathbf{x} = x_1x_2 \cdots x_n$ and $\mathbf{y} = y_1y_2 \cdots y_n$ in $GF(4)^n$ to be

$$\mathbf{x}{\star}\mathbf{y} = \sum_{i=1}^{n} \mathrm{Tr}(x_i \overline{y_i}).$$

Example 1.1. Let $\mathcal{G}_6$ be the $[6,3,4]$ *hexacode* whose generator matrix as a linear $GF(4)$-code is

$$\begin{bmatrix} 1 & 0 & 0 & 1 & \omega & \omega \\ 0 & 1 & 0 & \omega & 1 & \omega \\ 0 & 0 & 1 & \omega & \omega & 1 \end{bmatrix}.$$

This is also an additive code; thinking of $\mathcal{G}_6$ as an additive code, it has generator matrix

$$\begin{bmatrix} 1 & 0 & 0 & 1 & \omega & \omega \\ \omega & 0 & 0 & \omega & \overline{\omega} & \overline{\omega} \\ 0 & 1 & 0 & \omega & 1 & \omega \\ 0 & \omega & 0 & \overline{\omega} & \omega & \overline{\omega} \\ 0 & 0 & 1 & \omega & \omega & 1 \\ 0 & 0 & \omega & \overline{\omega} & \overline{\omega} & \omega \end{bmatrix}.$$

If $\mathcal{C}$ is an additive code, its *dual*, denoted $\mathcal{C}^\perp$, is the additive code $\{\mathbf{x} \in GF(4)^n \mid \mathbf{x}{\star}\mathbf{c} = 0 \text{ for all } \mathbf{c} \in \mathcal{C}\}$. If $\mathcal{C}$ is an $(n, 2^k)$ code, then $\mathcal{C}^\perp$ is an $(n, 2^{2n-k})$ code. As usual, $\mathcal{C}$ is *self-orthogonal* if $\mathcal{C} \subseteq \mathcal{C}^\perp$ and *self-dual* if $\mathcal{C} = \mathcal{C}^\perp$. In particular, if $\mathcal{C}$ is self-dual, $\mathcal{C}$ is an $(n, 2^n)$ code. The code $\mathcal{G}_6$ in Example 1.1 is self-dual as an additive code. (Any $GF(4)$-linear code that is self-orthogonal under the Hermitian inner product is a self-orthogonal additive code under the trace inner product.)

As usual, the *weight* $\mathrm{wt}(\mathbf{c})$ of $\mathbf{c} \in \mathcal{C}$ is the number of nonzero components of $\mathbf{c}$. The minimum weight d of $\mathcal{C}$ is the smallest weight of any nonzero codeword in $\mathcal{C}$. If $\mathcal{C}$ is an $(n, 2^k)$ additive code of minimum weight d, $\mathcal{C}$ is called an $(n, 2^k, d)$ code. $\mathcal{C}$ is *Type II* if $\mathcal{C}$ is self-dual and all codewords have even weight; Type II codes of length n exist only if n is even [6]. If $\mathcal{C}$ is self-dual but some codeword has odd weight (in which case the code cannot be $GF(4)$-linear), the code is *Type I* (see [11, section 4.2]). There is a bound on the minimum weight of an additive self-dual code [11, Theorem 33]. If d_I, d_{II} is the minimum distance of an additive self-dual Type I, Type II code, respectively, of length $n > 1$, then

$$d_I \le \begin{cases} 2\left\lfloor \frac{n}{6} \right\rfloor + 1 & \text{if } n \equiv 0 \pmod 6 \\ 2\left\lfloor \frac{n}{6} \right\rfloor + 3 & \text{if } n \equiv 5 \pmod 6 \\ 2\left\lfloor \frac{n}{6} \right\rfloor + 2 & \text{otherwise} \end{cases} \tag{1}$$

$$d_{II} \le 2\left\lfloor \frac{n}{6} \right\rfloor + 2. \tag{2}$$

A code that meets the appropriate bound is called *extremal.* Note that (2) is the same as saying that $d = 2m + 2$ if $n = 6m$. Type II codes meeting the bound d_{II} have a unique weight enumerator. This property is not true for Type I codes. A self-dual (with respect to Hermitian inner product) linear code over $GF(4)$ also satisfies bound (2) and an extremal code is a $[6m, 3m, 2m + 2]$ code.

We say that two additive codes $\mathcal{C}_1$ and $\mathcal{C}_2$ are *equivalent* provided there is a map sending the codewords of $\mathcal{C}_1$ onto the codewords of $\mathcal{C}_2$ where the map consists of a permutation of coordinates followed by a scaling of coordinates by elements of $GF(4)$ followed by conjugation of some of the coordinates. Notice that permuting coordinates, scaling coordinates, and conjugating some coordinates of a self-orthogonal (or self-dual) code does not change self-orthogonality (or self-duality). The *automorphism group* of $\mathcal{C}$, denoted $\mathrm{Aut}(\mathcal{C})$, consists of all maps which permute coordinates, scale coordinates, and conjugate coordinates that send codewords of $\mathcal{C}$ to codewords of $\mathcal{C}$.

The purpose of this paper is to study designs in additive codes over $GF(4)$. There are two types of designs. One is a classical t-design with repeated blocks. For these we have an analogue of the Assmus-Mattson Theorem for additive codes over $GF(4)$. The other is a generalized t-design first introduced by Delsarte [4]. As an example, we consider the dodecacode, which is the unique additive even self-dual $(12, 2^{12}, 6)$ code. We show that there exists a 5-$(12, 6, 3)$ design with multiplicity 1 or 3 in the dodecacode. We also derive a new simple 3-design, i.e. a 3-$(11, 5, 4)$ design from the shortened dodecacode.

2 Assmus-Mattson Theorem

We recall the Assmus-Mattson Theorem which guarantees simple t-designs in some linear codes and their duals over $GF(q)$.

Theorem 2.1 ([1, 10]). *Let $\mathcal{C}$ be an $[n, k, d]$ linear code over $GF(q)$ with dual code $\mathcal{C}^\perp$. Suppose $\mathcal{C}^\perp$ has minimum weight $d^\perp$. If $q = 2$, let $w = n$; if $q > 2$, let w be the largest integer so that*

$$w - \lfloor \frac{w+q-2}{q-1} \rfloor < d.$$

Define $w^\perp$ analogously using $d^\perp$. Suppose that $\{A_i : 0 \le i \le n\}$ and $\{B_i : 0 \le i \le n\}$ are the weight distributions of $\mathcal{C}$ and $\mathcal{C}^\perp$, respectively. Let s be the number of i with $B_i \neq 0$ for $0 < i \le n - t$. Suppose $t < d$ and $s \le d - t$. Then the following hold.
(i) The vectors of weight i in $\mathcal{C}$ hold a t-design provided $A_i \neq 0$ and $d \le i \le w$.
(ii) The vectors of weight i in $\mathcal{C}^\perp$ hold a t-design provided $B_i \neq 0$ and $d^\perp \le i \le \min\{n - t, w^\perp\}$.

Remark 2.2. The condition on w and $w^\perp$ in Theorem 2.1 ensures that there cannot be repeated blocks. This arises from the linearity of the code. Since we consider additive codes over $GF(4)$, there is no restriction on w and $w^\perp$ as in Theorem 2.7.

If $\mathcal{C}$ is a self-dual additive $(n, 2^n, d)$ code, we can obtain a self-dual code of length $n - 1$ by a process we call *shortening.* We use this term even though it differs from the usual meaning as our notion can be regarded as a generalization. Let G be a generator matrix of $\mathcal{C}$. As $\mathcal{C}$ is self-dual, no column of G is all 0's. Choose any column of G, say the i^{th} one. The entries in column i can be any of 0, 1, ω, or $\overline{\omega}$. By row reducing G (to obtain another generator matrix we call G again), we can make all the entries in column i equal to 0 except for one or two entries; if two they would be two of the three values 1, ω, or $\overline{\omega}$. The *shortened code of $\mathcal{C}$ on coordinate i*, denoted $\mathcal{C}'$ is the code with generator matrix G' obtained from G by eliminating one row of G with a nonzero entry in column i and then eliminating column i. [If there is only one nonzero entry in column i of G, then $\mathcal{C}'$ is $\mathcal{C}$ shortened in the usual sense.] Clearly $\mathcal{C}'$ is a self-dual additive $(n - 1, 2^{n-1}, d')$ code with $d' \ge d - 1$ [5, 6].

Let $\mathcal{C}$ be an additive $(n, 2^k, d)$ code over $GF(4)$. Let T be a set of t coordinate positions and I the remaining $n - t$ coordinate positions. Let $\mathcal{C}(I)$ be the code obtained from C by removing these t coordinate positions. Consider now the subcode D of $C^\perp$, which is 0 on all the positions in T. Let $D(I)$ be the code of length $n-t$ gotten by removing the t coordinate positions from D.

Remark 2.3. If $T=\{i\}$, then $D(I)$ is a shortened code of $C^\perp$ on coordinate i if $C^\perp$ has a generator matrix $G^\perp$ with only one nonzero element on i^{th} position. Otherwise, $D(I)$ is obtained by deleting the two rows of $G^\perp$ which have distinct nonzero elements on i^{th} position and eliminating the i^{th} column of $G^\perp$.

The following lemmas will be used to prove Theorem 2.7.

Lemma 2.4. *(i) $C(I)^\perp = D(I)$; moreover*
(ii) if $t<d$, then $dim(C(I)) = k$ and $dim(D(I)) = 2n-2t-k$.

Proof. Clearly $D(I) \subset C(I)^\perp$. Take any vector x' in $C(I)^\perp$. It is of length $n-t$ and if we extend it to a length n vector x by adjoining t zeros, then x certainly is in D. Hence x' should be in $D(I)$. This proves (i).

To prove (ii), suppose that we remove $t(<d)$ coordinate positions of C. If the dimension of $C(I)$ changes, this means that there exist $k'(<k)$ rows of $C(I)$ whose linear combination is a zero vector. Since these k' rows of $C(I)$ come from k' rows of C, we would have a vector in C of weight t less than d, which is absurd. Thus $\dim(C(I)) = k$ and hence $\dim(D(I)) = 2n-2t-k$. □

The following MacWilliams identities are known for additive codes over $GF(4)$ (see [11]).

Lemma 2.5. *Let C be an additive $(n, 2^k)$ code over $GF(4)$ and $C^\perp$ be the dual $(n, 2^{2n-k})$ code. Then their weight enumerators are related as follows.*

$$W_C(x,y) = \frac{1}{|C^\perp|} W_{C^\perp}(x+3y, x-y).$$

We give a uniqueness theorem which can be derived from Lemma 2.5.

Lemma 2.6. *If A_i's are the vectors of weight i in C and B_i's the vectors of weight i in $C^\perp$, then there exists a unique solution of A_i's and B_i's if there are at most s unknown values of A_i's and $B_1, \cdots, B_{s-1}$ are known.*

We now give an analogue of the Assmus-Mattson Theorem for additive codes over $GF(4)$ which gives designs with possibly repeated blocks. The original Assmus-Mattson Theorem gives simple designs.

Theorem 2.7. *Let C be an additive $(n, 2^k)$ code over $GF(4)$ with minimum weight d. Let $C^\perp$ be its dual $(n, 2^{n-k})$ code with minimum weight d'. Let $0<t<d$. Let s be the number of weights $B_i \neq 0$ in $C^\perp$ where $0<i\leq n-t$. Suppose that $s \leq d-t$. Then the following hold.*
(i) For each weight u $(d \leq u \leq n)$, the set of supports of codewords of weight u in C holds a t-design with possibly repeated blocks.
(ii) The set of supports of vectors of weight w in $C^\perp$ where $B_w \neq 0$ and $d' \leq w \leq n-t$ hold a t-design with possibly repeated blocks.
(iii) The supports of minimum weight vectors are either simple blocks or have repetition number 3.

Proof. Let T be a set of t coordinate positions and I the remaining $n-t$ coordinate positions. Then it follows from Lemma 2.4 that $C(I)^{\perp} = D(I)$ with $D(I)$ an $(n-t, 2^{2n-2t-k})$ code.

$(*)$ Let B'_i be the number of vectors of weight i in $D(I)$. Then $B'_i \leq B_i$, and if s' is the number of i so that $B'_i \neq 0$, then $s' \leq s$. Since we know by hypothesis $s \leq d-t$, we see $s' \leq d-t$; hence s' is less than or equal to the minimum weight of $C(I)$. Now we can apply Lemma 2.5 to the codes $C(I)$ and $D(I)$ and we obtain a unique solution to their weight distributions. This solution just depends on the dimensions, k and $2n-2k-t$ of the two codes. Thus it is the same for any set of t-coordinates chosen.

Suppose $u = d$. Let E_1 be the family of supports of codewords of C of weight d. These supports may be repeated. The number of blocks(=supports) in E_1 containing a given t-subset T is the number of vectors in $C(I)$ of weight $d-t$. By $(*)$, this number is independent of the choice of T. Hence the set of supports of codewords of C of weight d holds a t-design.

The rest of the proof of (i) is the same as that of the Assmus-Mattson theorem([10]) except that blocks can be repeated. So we omit the details.

$(**)$ We prove (ii). Let $d' \leq w \leq n-t$. Let F be the family of supports of codewords of weight w in $C^{\perp}$. Again supports can be repeated. Consider the family F' consisting of the complements of the sets in F. For any t-subset T of n coordinate positions, it follows that the number of blocks of F' containing T is equal to the number of vectors of weight w in $D(I)$. By the previous argument $(*)$ this number does not depend on the choice of T. Therefore, F' is a t-design and hence F is also a t-design since the complements of t-design with possibly repeated blocks give another t-design with the same number of repeated blocks.

For (iii), we first show that there is no block with multiplicity 2. Suppose that there are two codewords $\mathbf{u}$ and $\mathbf{v}$ of weight d with the same supports. Each nonzero component of $\mathbf{u}$ must be different from the corresponding component of $\mathbf{v}$ (otherwise we will have a codeword $\mathbf{u}+\mathbf{v}$ whose weight is less than d). Hence the sum $\mathbf{u}+\mathbf{v}$ is of weight d with the same support as $\mathbf{u}$ and $\mathbf{v}$. If there are 3 vectors of weight d with the same support, there cannot be more as in any given position all three of the non-zero values occur. ☐

Remark 2.8. The λ for these designs is computed by the usual formula from the weight distribution.

Corollary 2.9. *Let C be an extremal additive even self-dual $(n = 6m, 2^n)$ code over $GF(4)$ with minimum weight $d = 2m+2 \geq 6$. Then the vectors of each weight w in C where $A_w \neq 0$ and $d \leq w \leq 6m$ hold a 5-design with possibly repeated blocks.*

Proof. By (2) $2m+2$ is the minimum weight of an extremal even code of length $6m$. We note that the value s in Theorem 2.7 is $\{(6m-4)-(2m+2)\}/2 = 2m-3$, which equals $d-t = 2m+2-5 = 2m-3$. Hence this corollary follows immediately. ☐

Example 2.10. There exists a unique extremal additive even self-dual $(12, 2^{12}, 6)$ code up to equivalence [5]. It is called the *dodecacode* (see Table 1). By Corollary 2.9, the set of supports of vectors of weight 6 hold a 5-design with possibly repeated blocks. There are two kinds of blocks. One is a simple block and the other has multiplicity 3. We remark that there are 18 codewords of weight 6 whose supports form blocks repeated 3 times. Among them, exactly 12 codewords are such that their scalar multiples are also codewords. In the dodecacode, $A_6 = 396$. So there are 396-18=378 codewords of weight 6 whose supports are simple blocks. Note that $396 = \lambda_5 \binom{12}{3} / \binom{6}{3}$, whence $\lambda_5 = 3$. Allowing repeated blocks we obtain a 5-$(12, 6, \lambda_5 = 3)$ design in the dodecacode. Each 5 set is either contained in one block repeated 3 times or in 3 distinct blocks. It is the first example of a design of these parameters with multiplicity 1 or 3. Furthermore we see that vectors of other weights 8, 10, and 12 hold 5-designs with possibly repeated blocks with $\lambda = 105, 630$, and 234, respectively. See Table 2 for the weight distribution of the dodecacode.

Example 2.11. It was shown [5] that there exists a unique additive self-dual $(11, 2^{11}, 5)$ code up to equivalence. This is the shortened dodecacode See Table 1 for one such obtained by eliminating the last coordinate of the generator matrix QC_12 and its last row. We have shown by Magma that the vectors of weight 5 in the shortened dodecacode hold a simple 3-$(11, 5, 4)$ design. It can be seen in Corollary 3.3 that the minimum weight vectors always give simple blocks. We also get simple 2-$(11, 5, 12)$ and 1-$(11, 5, 30)$ designs. Also, the vectors of weight 6 hold a 4-(11,6,9) design with repeated blocks. See Table 2 for the weight distribution of the shortened dodecacode.

When we consider a punctured dodecacode on the last column of QC_12, we have 3-$(11, 5, 12)$, 2-$(11, 5, 36)$ and 1-$(11, 5, 90)$ designs with repeated blocks from weight 5 vectors and the same 4-$(11, 6, 9)$ design with repeated blocks from weight 6 vectors.

These examples lead to the following theorem.

Theorem 2.12. *There exists a 5-*$(12, 6, 3)$ *design with multiplicity 1 or 3 in the dodecacode. There exists a simple 3-*$(11, 5, 4)$ *design in the shortened dodecacode which is not equivalent to the 3-design obtained from the ternary Golay* $[11, 6, 5]$ *code with the same parameters.*

Proof. For the second claim, we note that both designs have 66 blocks and are 3-$(11, 5, 4)$ designs. So we compare automorphims of both designs. It is well known that the ternary Golay $[11, 6, 5]$ code has an automorphism which is a cycle of length 11 while the dodecacode does not have such a cycle. The automorphism group of the shortened dodecacode has order 18(see Table 2). ☐

Example 2.13. The existence of an extremal additive even self-dual $(24, 2^{24}, 10)$ code over $GF(4)$ is a long standing open question. If there exists such a code, it contains 5-designs for each even weight 10 to 24. There is no $GF(4)$ linear $[24, 12, 10]$ code [9].

3 Generalized t-designs

We give some definitions about generalized t-designs from [4]. Let F be a finite alphabet of size q (we are interested in $F = GF(4)$ with $q = 4$). Let $G = F^n$, the set of n-tuples of F.

An element a of G is said to be *covered* by an element b of G if each nonzero component a_i of a is equal to the corresponding component b_i of b; we denote this by $a \leq b$. For example, if $F = GF(4) = \{0, 1, \omega, \overline{\omega}\}$, then $a = (1, 1, \omega, 0)$ is covered by $b = (1, 1, \omega, \overline{\omega})$.

A subset S of G is called a *generalized t-design of type* $q - 1$, with parameters t-$(n, k, \mu_t), 0 \leq t \leq k \leq n, \mu_t \geq 1$, if the following two conditions are satisfied;

Table 1: Generator Matrices of Dodecacode and Shortened Dodecacode

$$G(QC\_12)=\begin{bmatrix}0&0&0&0&0&0&1&1&1&1&1&1\\0&0&0&0&0&0&\omega&\omega&\omega&\omega&\omega&\omega\\1&1&1&1&1&1&0&0&0&0&0&0\\\omega&\omega&\omega&\omega&\omega&\omega&0&0&0&0&0&0\\0&0&0&1&\omega&\overline{\omega}&0&0&0&1&\omega&\overline{\omega}\\0&0&0&\omega&\overline{\omega}&1&0&0&0&\omega&\overline{\omega}&1\\1&\overline{\omega}&\omega&0&0&0&1&\overline{\omega}&\omega&0&0&0\\\omega&1&\overline{\omega}&0&0&0&\omega&1&\overline{\omega}&0&0&0\\0&0&0&1&\overline{\omega}&\omega&\omega&\overline{\omega}&1&0&0&0\\0&0&0&\omega&1&\overline{\omega}&1&\omega&\overline{\omega}&0&0&0\\1&\omega&\overline{\omega}&0&0&0&0&0&0&\overline{\omega}&\omega&1\\\overline{\omega}&1&\omega&0&0&0&0&0&0&1&\overline{\omega}&\omega\end{bmatrix},G((QC\_12)')=\begin{bmatrix}0&0&0&0&0&0&1&1&1&1&1\\\overline{\omega}&1&\omega&0&0&0&\omega&\omega&\omega&\overline{\omega}&1\\1&1&1&1&1&1&0&0&0&0&0\\\omega&\omega&\omega&\omega&\omega&\omega&0&0&0&0&0\\\overline{\omega}&1&\omega&1&\omega&\overline{\omega}&1&1&1&1&0\\0&0&0&\omega&\overline{\omega}&1&1&1&1&\overline{\omega}&\omega\\1&\overline{\omega}&\omega&0&0&0&1&\overline{\omega}&\omega&0&0\\\omega&1&\overline{\omega}&0&0&0&\omega&1&\overline{\omega}&0&0\\0&0&0&1&\overline{\omega}&\omega&\omega&\overline{\omega}&1&0&0\\0&0&0&\omega&1&\overline{\omega}&1&\omega&\overline{\omega}&0&0\\1&\omega&\overline{\omega}&0&0&0&1&1&1&\omega&\overline{\omega}\end{bmatrix}.$$

Table 2: Weight Distribution and Group Order of QC_12 and $(QC\_12)'$

Code	n	d	$\lvert \mathrm{Aut}(C) \rvert$	A_0	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}	A_{11}	A_{12}
QC_12	12	6	648	1	0	0	396	0	1485	0	1980	0	234
$(QC\_12)'$	11	5	18	1	0	66	198	330	495	550	330	78	

(i) all elements of S have the same weight k,
(ii) each element of weight t in G is covered by a constant number μ_t of elements of S.
If a subset S of G holds a generalized t-design of type $q-1$, then it holds a generalized $t-1$-design of type $q-1$ [4].

In the binary case($q=2$), this design is the same as a classical t-design without repeated blocks.

For a given code $\mathcal{C}$ of length n and for an element e of $G=F^n$, we denote by $\mu(i,e)$ the number of codewords of weight i that cover e. Trivially, if $i < wt(e)$, then $\mu(i,e)=0$.

The Delsarte theorem for any finite alphabet is as follows.

Theorem 3.1 **([4]).** *Let $\mathcal{C}$ be a q-ary code of dual distance d'. Let t be an integer, $1 \le t \le d'$, such that the number of weights of $\mathcal{C}$ that are at least equal to t is at most equal to $d'-t$. Then each set of codewords of a given weight $\geq t$, is a generalized t-design of type $q-1$.*

Now we obtain generalized t-designs from additive codes over $GF(4)$.

Corollary 3.2 **([7]).** *Let $\mathcal{C}$ be an extremal even additive self-dual code over $GF(4)$ of length $n=6m$. Then the set of codewords of weight w in $\mathcal{C}$ with $A_w \neq 0$ forms a generalized 2-design of type 3.*

Proof. We observe that since $\mathcal{C}$ is extremal and even, its minimum weight is $2m+2$ which is also d'. As there are at most $2m$ non-zero weights in $\mathcal{C}$, the corollary follows. ☐

Corollary 3.3. *Let $\mathcal{C}$ be extremal, even, additive self-dual code over $GF(4)$ of length $n=6m$ with A_i its weight distribution. Let $\mathcal{C}'$ be a shortened code on any coordinate i with a_i its weight distribution. Then*

$$a_{2i+1}=\frac{A_{2i+2}}{3} \textit{ for } i=1,2,\cdots,m.$$

In particular, this implies that vectors of odd minimum weight in the shortened code are the only vectors with their support.

Proof. Since codewords of weight $w = 2i + 2$ in C with $A_w \neq 0$ hold a generalized 2-design of type 3 by Corollary 3.2, these vectors also hold a generalized 1-design of type 3. Thus for any given coordinate j there are exactly two non-zero elements in $GF(4)$, say 1 and ω, in the generator matrix G before we shorten C on the coordinate j. The shortened code C' obtained by eliminating one row of G with a nonzero entry, say ω, in column j and then eliminating column j has odd weight vectors of weight $2i + 1$ such that $a_{2i+1} = A_{2i+2}/3$ for $i = 1, 2, \cdots, m$. This is because we have excluded two other vectors of weight $2i + 2$ in C having nonzero entries distinct from ω. This completes the proof. □

Example 3.4. Here we compute $\mu(i, e)$ for the dodecacode. Let e be any vector of weight 2 in $G = (GF(4)^{12}$. Since codewords of nonzero weight i hold a generalized 2-design by Theorem 3.2, $\mu(i, e)$ is independent of the choice of e. We have $\mu(6, e) = 10, \mu(8, e) = 70, \mu(10, e) = 150$, and $\mu(12, e) = 26$. These numbers are obtained by dividing the λ for the 2-design with possibly repeated blocks held by these vectors by 9.

Example 3.5. We note that the vectors of weight 7 in the $(17, 2^{17}, 7)$ code shortened from the unique linear $[18, 9, 8]$ code S_{18} [8] contains simple 4,3, and 2 designs by Corollary 3.3 and the fact that S_{18} is linear. The 4-design is a 4-$(17, 7, 6)$ design.

References

[1] E. F. Assmus, Jr. and H. F. Mattson, Jr., "New 5-designs," J. Combin. Theory **6**(1969), pp. 122-151.

[2] C. Bachoc and P. Gaborit, "On extremal additive $GF(4)$-codes of lengths 10 to 18," To appear in J. Théorie Nombres Bordeaux.

[3] A. R. Calderbank, E. M. Rains, P. W. Shor, and N. J. A. Sloane, "Quantum error correction via codes over $GF(4)$," *IEEE Trans. Inform. Theory* **IT-44** (1998), pp. 1369–1387.

[4] Delsarte, "Four fundamental parameters of a code and their combinatorial significance," Inform. and Control **23**(1973), pp. 407–438,"

[5] P. Gaborit, W. C. Huffman, J.-L. Kim, and V. Pless, "On the classification of extremal additive codes over $GF(4)$," *Proceedings of the 37th Allerton Conference on Communication, Control, and Computing,* UIUC Sep. 1999, pp. 535–544.

[6] ——, "On additive $GF(4)$ codes," to appear in *Proceedings of the DIMACS Workshop on Codes and Association Schemes,* ed. A. Barg and S. Litsyn, AMS, Nov. 1999.

[7] G. Höhn, "Self-dual codes over the Kleinian four group," preprint, 1996. An updated version in http://xxx.lanl.gov/(math.CO/0005266).

[8] W. C. Huffman, "Characterization of quaternary extremal codes of lengths 18 and 20," *IEEE Trans. Inform. Theory* **IT-43** (1997), pp. 1613–1616.

[9] C. W. H. Lam and V. Pless, "There is no (24,12,10) self-dual quaternary code," *IEEE Trans. Inform. Theory,* vol. 36, pp. 1153–1156, Sept. 1990.

[10] V. S. Pless, W. C. Huffman and R. A. Brualdi, " An introduction to algebraic codes," in *Handbook of Coding Theory,* ed. V. S. Pless and W. C. Huffman. Amsterdam: Elsevier, 1998, pp. 3–139.

[11] E. M. Rains and N. J. A. Sloane, "Self-dual codes," in *Handbook of Coding Theory,* ed. V. S. Pless and W. C. Huffman. Amsterdam: Elsevier, 1998, pp. 177–294.

A Critical Look at Self-Dual Codes*

Judy L. Walker

Department of Mathematics and Statistics
University of Nebraska
Lincoln, NE 68588-0323
jwalker@math.unl.edu

Abstract

We investigate self-dual codes from a structural point of view. In particular, we study properties of critical indecomposable codes which appear in the spectrum of a self-dual code. As an application of the results we obtain, we revisit the study of self-dual codes of dimension at most 10.

1 Introduction

In the late 1950's, Slepian [4] became the first to take an abstract approach to the study of error-correcting codes. He introduced a structure theory for binary linear codes, developing in particular the idea of an *indecomposable code*; that is, a code which is not isomorphic to a nontrivial direct sum of two other codes. He proved two important results in this direction: First, every code is isomorphic to a unique sum of indecomposable codes. Second, for a given length and dimension, there is an indecomposable code which achieves the highest possible minimum distance.

The problem with indecomposable codes is that there are simply too many of them. A code is indecomposable of and only if it is not equivalent to a code which has a generator matrix which is block diagonal with at least two blocks. Thus, if C is any indecomposable code, then adding any column onto C yields a new indecomposable code of the same dimension but length one more than the length of C.

The major breakthrough in this area came in the late 1990's when Assmus ([1]) introduced the notion of *critical indecomposable* codes. The idea is that these codes are indecomposable codes with no "extra" columns tacked on. The notion of critical indecomposable codes appears to be very promising. In fact, Assmus shows that there is a "quasi-canonical" form for the generator matrix of such a code. Further, Assmus gives a recursive construction for all critical indecomposable codes.

A critical indecomposable code which can be obtained by puncturing an indecomposable code at one or more columns is said to be in the *spectrum* of that indecomposable code. In this paper, we investigate self-dual codes by considering properties of critical indecomposable codes which appear in the spectrum of a self-dual code. Recall that a code $C \subset \mathbb{F}_q^n$ is called *self-dual* if $C = C^\perp$, where $C^\perp := \{x \in \mathbb{F}_q^n \,|\, x \cdot c = 0 \text{ for every } c \in C\}$. These codes are known to have many remarkable properties, and much work has been done toward understanding them. In particular, considerable effort has been devoted to

---

*This research is supported in part by NSF Grants DMS-0071008 and DMS-0071011.

finding a complete enumeration of self-dual binary codes. Though the enumeration is now known through dimensions at least 16, the first paper on the subject ([3]) dealt with self-dual codes of dimension at most 10, a topic we revisit in this paper.

2 Critical Indecomposable Codes

We begin by recalling some definitions and results from [1] and [5]. In what follows, all codes are assumed to be binary, and we will write $\mathbb{F}$ to mean the field with two elements.

Definition 2.1. Let C' and C'' be linear codes of lengths m and n respectively. The *direct sum* of C' and C'' is the set $C := C' \oplus C''$ of all vectors $\mathbf{c} = (c_1, \ldots, c_{m+n} \in \mathbb{F}^{m+n}$ such that $(c_1, \ldots, c_m) \in C'$ and $(c_{m+1}, \ldots, c_{m+n}) \in C''$. Two codes are *isomorphic* if one can be obtained from the other through a permutation of coordinates. A code C is called *indecomposable* if it is not isomorphic to $C' \oplus C''$ for any nonzero linear codes C' and C''. A code which is not indecomposable is called *decomposable*.

Let C be an indecomposable code of length n and dimension k and let G be a generator matrix for C. For $1 \leq i \leq n$, define $\phi_i(C)$ to be the code which is generated by the rows of the matrix obtained from G by omitting the ith column. (This process is known as *puncturing* C at the ith column and is clearly independent of the choice of the generator matrix for C.) Then $\phi_i(C)$ will be a code of length $n-1$ and dimension either k or $k-1$, and it may or may not be indecomposable.

Definition 2.2. Let C be an indecomposable code of length n. We say the ith column of C is a *critical column* of C if either $\phi_i(C)$ has dimension $k-1$ or $\phi_i(C)$ is decomposable. We say C is a *critical indecomposable code* if every column of C is critical.

It is plain to see that $\mathbb{F}$ is the only critical indecomposable code of dimension 1. Further, for $k \geq 2$, there is only one code of dimension k and length $k+1$, and it is critical indecomposable. We will call this code $C_{k+1,k}$. There is also (see [5]) a unique critical indecomposable code $C_{2k-2,k}$ of length $2k-2$ and dimension k for each $k \geq 4$, and this is the longest critical indecomposable code of dimension k. Moreover, Assmus gives a recursive construction for all critical indecomposable codes.

The construction takes as input a partition $\pi = (x_1, \ldots, x_r, x')$ of the intended length n of the code and an *auxiliary code* A, which is indecomposable of length $s := r + x'$ and minimum distance at least 3. We require that $x_i \geq 2$ for $i = 1, \ldots, r$, and we assume without loss of generality that it is ordered so that $x_1 \geq x_2, \geq \cdots \geq x_r$. Further, we require that the last x' columns of A be critical. Then a generator matrix for a critical indecomposable code can be constructed from π and A as follows:

- For $1 \leq i \leq r$, let G_i be the $(x_i - 1) \times x_i$ matrix which consists of a column of 1's followed by the identity matrix of size $x_i - 1$. Note that if $x_i = 2$, then G_i is a generator matrix for the code whose only nonzero element is the vector $(1, 1)$; otherwise G_i generates the code $C_{x_i, x_i - 1}$ described above.

- Let $l = \dim A$ and fix a generator matrix G_A for A. For $1 \leq i \leq r$, let L_i be the $l \times x_i$ matrix whose first column is the ith column of G_A and whose other entries are all 0. Also let $L_{r+1}, \ldots, L_{r+x'}$ be the last x' columns of G_A.

- Set

$$G := \begin{pmatrix} G_1 & & & & & \\ & \ddots & & & & \\ & & G_r & & & \\ L_1 & \dots & L_r & L_{r+1} & \dots & L_{r+x'} \end{pmatrix}, \qquad (*)$$

where all blank spaces are assumed to be filled in with zeros.

The matrix G is a generator matrix for a critical indecomposable code of length n and dimension $k := n - s + l$. The relatively straight-forward proof is given in [1], and we will not reproduce it here. A generator matrix of this form is called *quasi-canonical* because the first $k - l$ rows are uniquely determined by π and the last l are determined by A.

It is also true (see [1] or [5]) that every critical indecomposable code is equivalent to a code with a generator matrix of the form $(*)$, and from this generator matrix, we may recover a partition and auxiliary code. Thus, given any critical indecomposable code C, it makes sense to refer to the partition and auxiliary code used to construct C.

In [5], we give a complete enumeration of critical indecomposable codes of dimension at most 10. This enumeration is given in terms of the partitions and auxiliary codes used in the constructions.

Finally, we present some notational conventions used in the remainder of the paper. For a matrix M, we will write M^t for the transpose of M. We will write $(\mathbf{w}_1|\mathbf{w}_2)$ to mean the vector formed by concatenating the vectors $\mathbf{w}_1$ and $\mathbf{w}_2$. If D is an indecomposable code, we will write $\mathrm{Spec}(D)$ for the *spectrum* of D, i.e., the set of critical indecomposable codes which can be obtained by puncturing D at one or more columns. Also, we will often blur the distinction between equivalent codes. For example, when $C \in \mathrm{Spec}(D)$, we will say that D has a generator matrix of the form $[G|M]$ where G is a generator matrix for C, rather than that D is equivalent to a code with a generator matrix of this form.

3 Applications to Self-Dual Codes

We begin with some results about critical indecomposable codes with lengths at the two extremes of the range of possibilities.

Theorem 3.1. *For each $k \geq 4$, the unique critical indecomposable code $C_{2k-2,k}$ of length $2k - 2$ and dimension k is in the spectrum of a self-dual code if and only if k is even. Further, when k is even, there is a unique self-dual code D with $C_{2k-2,k} \in \mathrm{Spec}(D)$ and D is doubly-even (i.e., all words in D have weight divisible by 4) if and only if $k \equiv 0 \pmod 4$.*

Proof. The quasi-canonical generator matrix for the code $C_{2k-2,k}$ is

$$G := \begin{pmatrix} 11 & & \\ & \ddots & \\ & & 11 \\ 10 & \dots & 10 \end{pmatrix},$$

so if D is a self-dual code with $C_{2k-2,k} \in \mathrm{Spec}(D)$, then D must have a generator matrix of the form $[G|M]$, where M is a $k \times 2$ matrix with rows $m_1, \dots, m_k$. If k is odd, then all rows of G have even weight; thus all rows of M must have even weight, but we also must have $m_i \cdot m_k = 1$ for $1 \leq i \leq k-1$ and this is impossible. If k is even, the only possibility for the rows of M is $m_i = (1,1)$ for $1 \leq i \leq k-1$ and (without loss of generality) $m_k = (1,0)$. Finally, we see that the rows of $[G|M]$ all have weight divisible by 4 if and only if $k \equiv 0 \pmod 4$, which completes the proof of the theorem. $\square$

Our next task is to show that no critical indecomposable code of dimension k and length $k+1$ can appear in the spectrum of a self-dual code. First we prove a lemma which will be used both in the proof of this statement and later in the paper.

Lemma 3.2. *Let M be a $r \times c$ matrix and set $S := MM^t$. Then the rank of S is at most c. Further, if every row of M has even weight, then then rank of S is at most $c-1$.*

Proof. The matrix S represents the composition $\mathbb{F}^r \to \mathbb{F}^c \to \mathbb{F}^r$, where the map $\mathbb{F}^r \to \mathbb{F}^c$ is given by M^t and the map $F^c \to \mathbb{F}^r$ is given by M. Therefore, the rank of S is at most the column rank of M, which is at most c. If every row of M has even weight, then $(1, \ldots, 1)^t$ is in the null space of M. Thus, in this case the column rank of M is at most $c-1$. $\square$

Theorem 3.3. *For any $k \geq 2$, the unique critical indecomposable code $C_{k+1,k}$ of length $k+1$ and dimension k is not in the spectrum of any self-dual code.*

Proof. Suppose $C_{k+1,k} \in \operatorname{Spec}(D)$ for some self-dual code D. Then D has a generator matrix of the form $[1|I|M]$, where every row of the $k \times (k-1)$ matrix M has even weight and the inner product of any two distinct rows of M is 1. The matrix MM^t is then the $k \times k$ matrix which has 0's on the diagonal and 1's elsewhere. Such a matrix has rank either k or $k-1$ depending on whether k is even or odd. However, by Lemma 3.2, the rank of MM^t is at most $k-2$. This shows that M, and hence D, cannot exist. $\square$

In studying which of the other critical indecomposable codes can appear in the spectrum of a self-dual code, we will find the next definition useful.

Definition 3.4. Let $\pi := (x_1, \ldots, x_r, x')$ be a partition of the integer n and set $s := r + x'$. The maps $i_\pi : \mathbb{F}^s \to \mathbb{F}^n$ and $j_\pi : \mathbb{F}^s \to \mathbb{F}^n$ are defined by $i_\pi(a_1, \ldots, a_s) := (a_1^{x_1}, \ldots, a_r^{x_r}, a_{r+1}, \ldots, a_s)$ and $j_\pi(a_1, \ldots, a_s) := (a_1, 0^{x_1 - 1}, \ldots, a_r, 0^{x_r}, a_{r+1}, \ldots, a_s)$, where a^x is the vector of length x all of whose entries are a.

Lemma 3.5. [1] *Let C be a critical indecomposable code. Let π be the partition and let A be the auxiliary code used in the construction of C. Then $C^\perp = i_\pi(A^\perp)$.*

Proof. Let n be the length of C, k the dimension of C, s the length of A, and l the dimension of A. Then $n - k = s - l$, which means that $\dim(C^\perp) = \dim(A^\perp) = \dim(i_\pi(A^\perp))$. Thus, it is enough to show that $i_\pi(\mathbf{b}) \cdot \mathbf{f} = 0$ for every $\mathbf{b} \in A^\perp$ and every row $\mathbf{f}$ in a quasi-canonical generator matrix for C. Let $\mathbf{b} = (b_1, \ldots, b_s) \in A^\perp$. If $\mathbf{f}$ is one of the first $k - l$ rows of a quasi-canonical generator matrix for C, then $\mathbf{f}$ has weight 2 and support contained entirely within the block corresponding to some x_i, $1 \leq i \leq r$, where $\pi = (x_1, \ldots, x_r, x')$. Thus $i_\pi(\mathbf{b}) \cdot \mathbf{f} = b_i + b_i = 0$. On the other hand, if $\mathbf{f}$ is one of the last l rows, then $\mathbf{f} = j_\pi(\mathbf{a})$ for some $\mathbf{a} \in A$. Thus $i_\pi(\mathbf{b}) \cdot \mathbf{f} = 0$ since $\mathbf{b} \in A^\perp$. $\square$

Theorem 3.6. *Let C be a critical indecomposable code and suppose that C is in the spectrum of some self-dual code. Then $C^\perp \subseteq C$.*

Proof. Let D be a self-dual code with $C \in \operatorname{Spec}(D)$. Let n and k be the length and dimension of C so that D has length $2k$. We may assume that C is obtained by puncturing D at the last $2k - n$ columns. Let $\mathbf{z} = (z_1, \ldots, z_n) \in C^\perp$ and let $\mathbf{y} = (y_1, \ldots, y_{2k})$ be the vector of length $2k$ such that $y_i = z_i$ for $1 \le i \le n$ and $y_i = 0$ for $n+1 \le i \le 2k$. Let $\mathbf{d} = (d_1, \ldots, d_{2k})$ be any codeword in D. Then $\mathbf{y} \cdot \mathbf{d} = \sum y_i d_i = \sum z_i d_i = 0$ since $(d_1, \ldots, d_n) \in C$. Hence $\mathbf{y} \in D^\perp = D$, and so $\mathbf{z} \in C$. □

This theorem turns out to be quite powerful and will severely restrict which critical indecomposable codes can appear in the spectrum of a self-dual code. The crucial step in obtaining this restriction is given by the next lemma.

Lemma 3.7. *Let C be the critical indecomposable code constructed from the partition $\pi = (x_1, \ldots, x_r, x')$ and the auxiliary code A, and suppose C is in the spectrum of some self-dual code. Let $\mathbf{b}$ be any codeword of $A^\perp$. If the only codeword of A with support contained in* $\operatorname{supp}(\mathbf{b})$ *is* 0, *then* $\operatorname{supp}(\mathbf{b}) \subset \{1, \ldots, r\}$ *and x_i is even for every $i \in \operatorname{supp}(\mathbf{b})$. If* $\operatorname{supp}(\mathbf{b}) \not\subset \{1, \ldots, r\}$, *then there is a codeword $\mathbf{a} \in A$ with* $\operatorname{supp}(\mathbf{b}) \setminus \{1, \ldots, r\} \subset \operatorname{supp}(\mathbf{a}) \subset \operatorname{supp}(\mathbf{b})$.

Proof. Let $\mathbf{b} \in A^\perp$. Then $i_\pi(\mathbf{b}) \in C$, and so $i_\pi(\mathbf{b}) = \mathbf{e} + j_\pi(\mathbf{a})$ for some $\mathbf{e} := (\mathbf{e}_1 | \ldots | \mathbf{e}_r) \in E_1 \oplus \cdots \oplus E_r$ and some $\mathbf{a} := (a_1, \ldots, a_r, a_{r+1} \ldots, a_{r+x'}) \in A$. For $r+1 \le i \le r+x'$, we see that $b_i = a_i$, so the supports of $\mathbf{b}$ and $\mathbf{a}$ are identical in the last x' columns. For $1 \le i \le r$, if $a_i = 1$, then $b_i = 1$ since E_i has minimum distance 2. Thus, $\operatorname{supp}(\mathbf{a}) \subset \operatorname{supp}(\mathbf{b})$. In particular, if 0 is the only codeword of A with support contained in $\operatorname{supp}(\mathbf{b})$, then the last x' coordinates of $\mathbf{b}$ must be zero. Further, in that case, we must have the all-one vector contained in E_i for each $i \in \operatorname{supp}(\mathbf{b})$, which implies that x_i must be even for all $i \in \operatorname{supp}(\mathbf{b})$. □

We now use this lemma to find necessary conditions on the ingredients used to construct a critical indecomposable code which appears in the spectrum of a self-dual code.

Theorem 3.8. *Let $C = C(\pi, A)$ be the critical indecomposable code constructed using the partition $\pi = (x_1, \ldots, x_r, x')$ and the auxiliary code A, where the dimension of C is at most* 10. *If C is in the spectrum of a self-dual code, then $x' = 0$ and each x_i is even.*

Proof. All critical indecomposable codes of dimension at most 10 are listed in [5] according to the partitions and auxiliary codes used to construct them. The idea of this proof is to consider each auxiliary code in turn. To set up notation, let $s := r + x'$ and l be the length and dimension of the auxiliary code A.

If $l = 1$, then $A = A_{s,1}$ for some s and $A^\perp = C_{s,s-1}$. Since A has no critical columns, we already have $x' = 0$ and $r = s$. Further, since $i_\pi(A^\perp) = C^\perp \subseteq C$ by Lemma 3.5 and Theorem 3.6, we have that $i_\pi(\mathbf{f}_i) \in C$ where $\mathbf{f}_i$ is the ith row of the generator matrix $[I|1]$ for $C_{s,s-1}$. But $\operatorname{supp}(i_\pi(\mathbf{f}_i))$ is the set of coordinates in the blocks corresponding to x_i and x_s. This means that x_i and x_s are even by Lemma 3.7. Since $\mathbf{f}_i \in C_{s,s-1}$ for $1 \le i \le s-1$, we have that x_j is even for all j.

Now suppose $l = 2$. The shortest admissible code of dimension 2 has length 5, and so we have $s \ge 5$. Without loss of generality we may assume that the three nonzero codewords of A are $(1^u, 0^v, 1^w)$, $(0^u, 1^v, 1^w)$, and $(1^u, 1^v, 0^w)$ where u, v, and w satisfy $u \ge 2$, $v \ge 2$, and $u + v + w = s$. Let $u_0 = 2\lfloor u/2 \rfloor$, so that u_0 is the largest even integer which is not larger than u. Then $(1^{u_0}, 0^{u-u_0}, 0^v, 0^w) \in A^\perp$ and by Lemma 3.7, x_1, …, x_{u_0} must all be even. Since the vector $(0^{u-2}, 1^2, 0^v, 0^2)$ is also in $A^\perp$, we get that x_{u-1} and x_u are both even. Likewise, we can show that $x_{u+1}, \ldots, x_{u+v}$ are all even. If

$x' = 0$, a similar argument (using the vector $(1, 0^{u-1}, 1, 0^{v-1}, 1^w)$ if w is odd, or the vectors $(1, 0^{u-1}, 1, 0^{v-1}, 1^{w-1}, 0)$ and $(1, 0^{u-1}, 1, 0^{v-1}, 0, 1^{w-1})$ if w is even) shows that $x_{u+v+1}, \dots, x_{u+v+w}$ are all even. In summary, when $l = 2$ and $x' = 0$, each x_i must be even.

If $l = 2$ and $x' = 1$, then A must have a critical column, which we may assume is the last column of A. This means, in particular, that $w = 1$ in this situation. Then $(1, 0^{u-1}, 1, 0^{v-1}, 1) \in A^\perp$, and so by Lemma 3.7, we must have a vector $\mathbf{a} \in A$ with $1 \in \operatorname{supp}(\mathbf{a}) \subset \{1, u + 1, s\}$. Such an $\mathbf{a}$ does not exist, and so x' must be 0.

Next, suppose $l = 3$. There is no nice general description of the nonzero codewords in admissible codes of dimension 3. However, we still use the basic idea of finding explicit codewords in the dual code and applying Lemma 3.7. For example, consider the admissible code $A_{7,3;1}$, which has quasi-canonical generator matrix

$$\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}.$$

One sees immediately that $(1, 1, 0, 0, 0, 0, 0)$, $(0, 0, 1, 1, 0, 0, 0)$, and $(0, 0, 0, 0, 1, 1, 0)$ are all in $A_{7,3;1}^\perp$, and applying Lemma 3.7, we find that any partition used with $A_{7,3;1}$ must have x_i even for $1 \le i \le 6$. Taking any word of $A_{7,3;1}^\perp$ supported on the last column (for example, $(1, 0, 1, 0, 0, 1, 1)$), we find that x' must be 0 and x_7 must be even as well.

The arguments for the rest of the admissible codes are the same, using the codewords in the dual codes supplied by the following table.

A	relevant words in $A^\perp$
$A_{6,3;0}$	$(0,1,0,1,1,0)$, $(1,1,0,0,0,1)$, and $(0,0,1,1,0,1)$
$A^1_{7,3;0}$	$(0,1,0,1,1,0,0)$, $(1,1,0,0,0,1,0)$, $(0,0,1,1,0,1,0)$, and $(1,0,0,0,0,0,1)$
$A^2_{7,3;0}$	$(0,1,0,1,1,0,0)$, $(1,1,0,0,0,1,0)$, $(0,0,1,1,0,1,0)$, and $(1,0,0,1,0,0,1)$
$A^1_{8,3;1}$	$(1,1,0,0,0,0,0,0)$, $(0,1,1,0,0,0,0,0)$, $(0,0,0,1,1,0,0,0)$, $(0,0,0,0,0,1,1,0)$, and $(1,0,0,0,1,0,1,1)$
$A^2_{8,3;1}$	$(1,1,0,0,0,0,0,0)$, $(0,0,1,1,0,0,0,0)$, $(0,0,0,0,1,1,0,0)$, $(1,0,0,0,0,1,1,0)$, and $(0,0,1,0,0,0,1,1)$
$A_{7,3;2}$	$(1,1,0,0,0,0,0)$, $(0,0,1,1,0,0,0)$, $(0,0,0,1,1,1,0)$, and $(1,0,0,0,1,0,1)$
$A^1_{8,3;2}$	$(1,1,0,0,0,0,0,0)$, $(0,1,1,0,0,0,0,0)$, $(0,0,0,1,1,0,0,0)$, $(0,0,0,1,0,1,0,1)$, and $(1,0,0,0,0,1,1,0)$
$A^2_{8,3;2}$	$(1,1,0,0,0,0,0,0)$, $(0,0,1,1,0,0,0,0)$, $(0,0,0,0,1,1,0,0)$, and $(0,1,1,0,0,0,1,1)$
$A^1_{9,3;2}$	$(1,1,0,0,0,0,0,0,0)$, $(0,0,1,1,0,0,0,0,0)$, $(0,0,0,0,1,1,0,0)$, $(1,0,0,0,0,0,1,1,0)$, and $(0,0,0,0,1,0,1,0,1)$
$A^2_{9,3;2}$	$(1,1,0,0,0,0,0,0,0)$, $(0,1,1,0,0,0,0,0,0)$, $(0,0,0,1,1,0,0,0,0)$, $(0,0,0,0,0,1,1,0,0,0)$, $(1,0,0,0,0,0,1,1,0)$, and $(0,0,0,1,0,0,1,0,1)$
$A^3_{9,3;2}$	$(1,1,0,0,0,0,0,0,0)$, $(0,1,1,0,0,0,0,0,0)$, $(0,0,0,1,1,0,0,0,0)$, $(0,0,0,0,0,1,1,0,0)$, and $(0,0,1,1,0,0,0,1,1)$
$A^4_{9,3;2}$	$(1,1,0,0,0,0,0,0,0)$, $(0,0,1,1,0,0,0,0,0)$,$(0,0,0,0,1,1,0,0,0)$, $(0,0,0,0,0,1,1,0,0)$, and $(0,1,1,0,0,0,0,1,1)$
$A^1_{9,4;3}$	$(1,1,0,0,0,0,0,0,0)$, $(0,0,1,1,0,0,0,0,0)$, $(0,0,0,0,1,1,0,1,0)$, and $(0,1,1,0,0,0,1,1,1)$
$A^2_{9,4;3}$	$(1,1,0,0,0,0,0,0,0)$, $(0,0,1,1,0,0,0,0,0)$, $(0,0,0,0,1,1,0,0,0)$, and $(0,0,0,0,0,0,1,1,1)$

$\square$

4 Self-Dual Codes of Dimension At Most 10

In this section, we apply the results obtained above to revisit the enumeration of self-dual codes of dimension at most 10 given by Pless ([3]). It is hoped that these examples will demonstrate how one might hope to obtain new results in larger dimensions using these techniques. As before, we refer to auxiliary codes using the notations of [5].

We treat each dimension $k \leq 10$ in turn. The first five dimensions are easy: When $k = 1$, the code D_2 with generator matrix $[1, 1]$ is the unique self-dual code of dimension 1. When $k = 2$ or 3, the only critical indecomposable code is $C_{k+1,k}$, and so there are no indecomposable self-dual codes in these dimensions by Theorem 3.3. When $k = 4$, there are two critical indecomposable codes: $C_{5,4}$ and $C_{6,4}$. We know $C_{5,4}$ is not in the spectrum of any indecomposable self-dual code by Theorem 3.3, and by Theorem 3.1, $C_{6,4}$ is in the spectrum of a unique indecomposable self-dual code which we will call D_8. By the same theorem, we know that D_8 is in fact doubly-even. By Theorems 3.3, 3.7, and 3.1, we see that none of the three critical indecomposable codes of dimension 5 can be in the spectrum of any self-dual code, and so there are no indecomposable self-dual codes of dimension 5.

The case where $k = 6$ requires a bit more thought. By Theorem 3.8, the only critical indecomposable codes which might appear in the spectrum of an indecomposable code of dimension 6 are $C((4, 2^2), A_{3,1})$ and $C((2^5), A_{5,1})$, and by Theorem 3.1, this latter code is in the spectrum of a unique self-dual code which we can call D_{12}. In fact, this is the only indecomposable self-dual code of dimension 6. Indeed, any other one must have $C((4, 2^2), A_{3,1})$ in its spectrum. This means it would have a generator matrix of the form $[G|M]$ where G is the quasi-canonical generator matrix for $C((4, 2^2), A_{3,1})$ and M is a 6×4 matrix, chosen carefully so that each pair of rows of $[G|M]$ has inner product 0. Without loss of generality, then, M must be either

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}^t \quad \text{or} \quad \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}^t.$$

In either case, one gets a code equivalent to D_{12}.

For $k = 7$, we have that the only critical indecomposable codes which might appear in the spectrum of a self-dual code are $C_1 := C((4, 2^3), A_{4,1})$ and $C_2 := C((2^5), A_{5,2;1})$ by Theorems 3.8 and 3.1. Suppose first that C_1 is in the spectrum of a self-dual code. Then that self-dual code has a generator matrix of the form $[G|M]$ where G is the quasi-canonical generator matrix for C_1 and M is a 7×4 matrix satisfying

$$MM^t = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}.$$

By Lemma 3.2, the matrix MM^t must have rank at most 3, but a quick check shows that this matrix has rank 4. Thus C_1 cannot appear in the spectrum of any self-dual code. On the other hand, it is not hard to see that C_2 is in the spectrum of a unique self-dual code: any such code must have a generator matrix of the form $[G|M]$, where G is the quasi-canonical generator matrix for C_2 and the only possible choice for M (up to equivalence) is

$$M = \begin{pmatrix} 1 & 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 \end{pmatrix}^t .$$

We write D_{14} for this unique indecomposable self-dual code of dimension 7.

Things are only slightly more complicated for $k = 8$. Theorem 3.8 shows that the only critical indecomposable codes which might appear in the spectrum of a self-dual code are $C_1 := C((6,2^2), A_{3,1})$, $C_2 := C((4^2,2), A_{3,1})$, $C_3 := C((4,2^4), A_{5,1})$, $C_4 := C((2^6), A_{6,2;0})$, $C_5 := C((2^6), A_{6,2;1})$, and $C_6 := C((2^7), A_{7,1})$. Further, Theorem 3.1 shows that C_6 is in the spectrum of a unique (doubly-even) self-dual code D_{16}^1. We can rule out C_4 using Lemma 3.2 since the generator matrix of a self-dual code having C_4 in its spectrum would be $[G|M]$ where G is a quasi-canonical generator matrix for C_4 and M is a 8×4 matrix all of whose rows have even weight and satisfying

$$MM^t = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \end{pmatrix},$$

and this matrix has rank 4. It is easy to see that there is a unique self-dual code with C_3 in its spectrum: the generator matrix must be $[G|M]$ where G is the quasi-canonical generator matrix for C_3 and M is

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}^t$$

Further, this code is not equivalent to our code D_{16}^1, and so we call it D_{16}^2. At this point, one can use a brute force approach to show that these are the only two indecomposable self-dual codes of dimension 8 by constructing all the different ways that C_1, C_2, and C_5 could be extended to self-dual codes. Along the way, we discover that $\mathrm{Spec}(D_{16}^1) = \{C_1, C_6\}$ and $\mathrm{Spec}(D_{16}^2) = \{C_1, C_2, C_3, C_5\}$.

For $k = 9$, we can proceed in a similar fashion. Using Theorems 3.7 and 3.1, we find that $C_1 := C((6,2^3), A_{4,1})$, $C_2 := C((4^2,2^2), A_{4,1})$, $C_3 := C((4,2^4), A_{5,2;1})$, $C_4 := C((4,2^4), (A_{5,2;1})^{(1,5)})$, $C_5 := C((2^6), A_{6,3;0})$, $C_6 := C((4,2^5), A_{6,1})$, $C_7 := C((2^7), A_{7,2;0})$, $C_8 := C((2^7), A_{7,2;1}^1)$, and $C_9 := C((2^7), A_{7,2;1}^2)$ are the only critical indecomposable codes which could possibly appear in the spectrum of a self-dual code. Using Lemma 3.2, we can rule out C_1, C_2, C_6, and C_9. We can construct two inequivalent indecomposable self-dual codes D_{18}^1 and D_{18}^2 by appending

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 1 \end{pmatrix}^t \quad \text{and} \quad \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \end{pmatrix}^t$$

to the quasi-canonical generator matrices of C_3 and C_7, respectively. By using brute force to see what ways the various remaining critical indecomposable codes can extend, one finds that these are the only indecomposable self-dual codes of dimension 9. Along the way, we find that $\text{Spec}(D_{18}^1) = \{C_3, C_4, C_8\}$ and $\text{Spec}(D_{18}^2) = \{C_3, C_4, C_5, C_7\}$.

Finally, there are 16 critical indecomposable codes of dimension 10 that could occur in the spectrum of a self-dual code by Theorems 3.8 and 3.2. They are:

Name	Partition	Auxiliary Code	Name	Partition	Auxiliary Code
C_1	$(8, 2^2)$	$A_{3,1}$	C_2	$(6, 4, 2)$	$A_{3,1}$
C_3	(4^3)	$A_{3,1}$	C_4	$(6, 2^4)$	$A_{5,1}$
C_5	$(4^2, 2^3)$	$A_{5,1}$	C_6	$(4, 2^5)$	$A_{6,2;1}$
C_7	$(4, 2^5)$	$(A_{6,2;1})^{(1,4)}$	C_8	$(4, 2^5)$	$(A_{6,2;1})^{(1,6)}$
C_9	(2^7)	$A_{7,3;0}^1$	C_{10}	(2^7)	$A_{7,3;1}$
C_{11}	(2^7)	$A_{7,3;2}$	C_{12}	$(4, 2^6)$	$A_{7,1}$
C_{13}	(2^8)	$A_{8,2;0}^1$	C_{14}	(2^8)	$A_{8,2;1}^1$
C_{15}	(2^8)	$A_{8,2;1}^2$	C_{16}	(2^8)	$A_{9,1}$

From Theorem 3.1, we know that C_{16} is in the spectrum of a unique indecomposable (and doubly-even) self-dual code, D_{20}^1. One could use a brute force approach with the other 15 codes, but this seems rather lengthy. Alternatively, one can do a random attack, producing new self-dual codes and checking to see if the mass formula (see, for example, [3]) is satisfied after each one. Further, one can compute (by hand or using the computer algebra system MAGMA ([2]), for example) the spectrum of each indecomposable self-dual code constructed. We follow this latter approach. We see that $\text{Spec}(D_{20}^1) = \{C_1, C_{16}\}$. Next, we notice that the only way for $[G|M]$ to be a generator matrix for a self-dual code, if G is the quasi-canonical generator matrix for C_{12}, is if M is the matrix

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}^t .$$

The self-dual code D_{20}^2 produced in this way is obviously not equivalent to D_{20}^1 since $C_{12} \notin \text{Spec}(D_{20}^1)$. Further, we can compute $\text{Spec}(D_{20}^2) = \{C_1, C_2, C_4, C_8, C_{12}, C_{14}\}$. Trying twice more, we find that C_5 and C_{15} each extend uniquely, using the matrices

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{pmatrix}^t \quad \text{and} \quad \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 \end{pmatrix}^t ,$$

respectively. Calling the self-dual codes we get D_{20}^3 and D_{20}^4, we compute their spectra to be $\{C_1, C_2, C_3, C_5, C_6, C_7, C_8, C_9, C_{10}, C_{11}\}$ and $\{C_1, C_2, C_3, C_4, C_7, C_8, C_{10}, C_{11}, C_{13}, C_{15}\}$ respectively. By the mass formula, there are still more indecomposable self-dual codes of dimension 10. At this point, we must use brute force on one of our longer critical indecomposable codes. Trying C_{11}, we see that the only possible ways to extend this code to a self-dual code are to add one of the following matrices:

$$\begin{pmatrix} 1&1&1&1&1&0&0&1&\bar{x}&x+y \\ 1&1&1&1&1&0&0&0&x&\overline{x+y} \\ 0&0&1&1&0&1&0&x&1&y+z \\ 0&0&1&1&0&1&0&x&0&\overline{y+z} \\ 0&0&0&0&1&1&1&y&z&1 \\ 0&0&0&0&1&1&1&y&z&0 \end{pmatrix}^t, \begin{pmatrix} 1&1&0&0&0&0&1&1&x&z \\ 1&1&0&0&0&0&1&0&\bar{x}&\bar{z} \\ 0&0&1&1&0&1&0&1&y&\bar{x} \\ 0&0&1&1&0&1&0&0&y&\bar{x} \\ 0&0&0&0&1&1&1&1&\overline{x+y}&\overline{x+z} \\ 0&0&0&0&1&1&1&0&x+y&\overline{x+z} \end{pmatrix}^t,$$

$$\begin{pmatrix} 1&1&1&1&1&0&0&1&\overline{x+y}&x \\ 1&1&1&1&1&0&0&0&\overline{x+y}&\bar{x} \\ 1&1&0&0&0&0&1&x&1&z \\ 1&1&0&0&0&0&1&x&0&\bar{z} \\ 0&0&0&0&1&1&1&y&1&z \\ 0&0&0&0&1&1&1&y&0&z \end{pmatrix}^t, \text{ or } \begin{pmatrix} 1&1&1&1&1&0&0&1&\bar{y}&x \\ 1&1&1&1&1&0&0&0&y&x \\ 1&1&0&0&0&0&1&x&z&1 \\ 1&1&0&0&0&0&1&x&z&0 \\ 0&0&1&1&0&1&0&y&1&z \\ 0&0&1&1&0&1&0&y&0&x+z \end{pmatrix}^t,$$

where x, y, and z can be 0 or 1, and $\bar{a}$ means $1+a$. With the first or third of these matrices, we get D_{20}^3 no matter what choices we make for x, y, and z. We also get D_{20}^3 when $x=0$ with the second matrix (independent of y and z), and we get D_{20}^4 when $x=z$ (independent of y) with the fourth matrix. However, when we use the second matrix with $x=1$, we get a new code which we call D_{20}^5, and when we use the fourth matrix with $x \neq z$, we get another one, which we call D_{20}^6. We can now use the mass formula to show that these are all the possibilities. Finally, we compute $\text{Spec}(D_{20}^5) = \{C_{10}, C_{11}\}$ and $\text{Spec}(D_{20}^6) = \{C_1, C_2, C_3, C_6, C_9, C_{11}\}$.

References

[1] E. F. Assmus, Jr. The category of linear codes. *IEEE Transactions on Information Theory*, 44:612–629, 1998.

[2] W. Bosma and J. J. Cannon. *Handbook of Magma Functions.* Sydney, Australia, 1996.

[3] V. Pless. A classification of self-orthogonal codes over GF(2). *Discrete Math*, 3:209–246, 1972.

[4] D. Slepian. Some further theory of group codes. *Bell Syst. Tech. Journal*, 39:1219–1252, 1960.

[5] J. L. Walker. Constructing critical indecomposable codes. Preprint; May, 2000.

The derivatives of Boolean functions: some related problems in coding theory and cryptography

Pascale Charpin
INRIA, Domaine de Voluceau,
Rocquencourt, BP 105 - 78153,
Le Chesnay Cedex, FRANCE.
Pascale.Charpin@inria.fr

Extended abstract[1]

Most conventional cryptographic primitives are concerned by the properties of Boolean functions (block ciphers, stream ciphers, hash functions). It is well-known that the nonlinearity of Boolean functions, *i.e.* their Hamming distances to the set of affine functions, is a property of most interest. On the other hand, the *propagation characteristics* for a function f, describe the behaviors of their derivatives $D_af \,:\, x \mapsto f(x+a)+f(x)$.

These properties, considering associated codewords of Boolean functions, can be viewed in the ambient space of primitive binary codes. Nonlinearity is related to the weight enumerators of the cosets of the Reed-Muller code of order 1. Any derivative can be seen as a multiplication in a group algebra. The use of both representations (a function and its codeword) is efficient for the description of any corpus of functions, characterizing *good* or *weak* cryptographic functions.

Note that the finite field of order 2^m is considered as a vector-space over the field of order 2: $\{\mathbf{F}_{2^m},+\} \approx \mathbf{F}_2^m$. Let $f \in B_m$ denote any Boolean function of m variables, *i.e.* from $\mathbf{F}_2^m$ to $\mathbf{F}_2$. On the other hand, we consider $\{\mathbf{F}_{2^m},+\}$ as an ordered set, say $\{\ \alpha_1,\ldots,\alpha_{2^m}\ \}$. We then define the *associated codeword of* f:

$$\Omega_f = (\ f(\alpha_1),\ \ldots, f(\alpha_{2^m})\ )$$

This codeword can be viewed in the ambient space of binary primitive codes, *the field algebra* $\mathcal{A} = \mathbf{F}_2[\mathbf{F}_{2^m},+,\times]$ (cf. an extensive study in [3]):

$$\Omega_f = \sum_{i=1}^{2^m} f(\alpha_i) X^{\alpha_i}\ .$$

A function $f \in B_m$ is said to be *balanced* if its Fourier transform equals 0 or, equivalently, if the weight of Ω_f equals 2^{m-1}.

The *nonlinearity* of f is the minimum weight of the coset of the Reed-Muller code of order one $\Omega_f + R(1,m)$. The concept of "derivative" can be seen as a multiplication in $\mathcal{A}$, as follows: for any $a \in \mathbf{F}_{2^m}$, the associated codeword of the function D_af is $\Omega_{D_af} = (X^a + 1)\Omega_f$. This observation is particularly interesting for the study of the functions which have linear structures, *i.e.* derivatives which are constant.

---

[1] This paper is an extended abstract of my talk, where I explained a part of a joint work with A. Canteaut, C. Carlet and C. Fontaine [1, 2].

We say that a Boolean function is *almost optimal* when its nonlinearity is "high" ; we define this corpus by taking in account the known bounds on the covering radius of $R(1, m)$.

Definition. Any coset $\mathrm{y} + R(1, m)$ is called an *almost optimal* coset when its minimum weight δ_{y} satisfies :

- $2^{m-1} - 2^{\frac{m}{2}} \leq \delta_{\mathrm{y}} \leq 2^{m-1} - 2^{\frac{m}{2}-1}$ for even m. It is *three-valued almost optimal* when it has three weights only. These are: 2^{m-1} and $2^{m-1} \pm 2^{\frac{m}{2}}$.
- $2^{m-1} - 2^{\frac{m-1}{2}} \leq \delta_{\mathrm{y}} < 2^{m-1} - 2^{\frac{m}{2}-1}$ for odd m. It is *three-valued almost optimal* when it has three weights only. These are: 2^{m-1} and $2^{m-1} \pm 2^{\frac{m-1}{2}}$.

Any Boolean function f is almost optimal if and only if the coset $\Omega_f + R(1, m)$ is almost optimal ; moreover f is three-valued if and only if $\Omega_f + R(1, m)$ has three weights only.

We give a necessary condition for a coset $\mathrm{y} + R(1, m)$ to be almost optimal. This condition leads to a property of cubic functions: *if a cubic function is almost optimal then it has at least λ derivatives which are balanced where $\lambda = 2^{m-2} - 1$ for even m and $\lambda = 2^{m-1} - 1$ for odd m.*

As an example, another consequence of this necessary condition is the description of the set of almost optimal functions which have linear structure. We obtain for odd m a full description.

Proposition. Let m be an odd integer, $m \geq 3$, $f \in \mathcal{B}_m$ and $e \in \mathbf{F}_{2^m}$ such that D_ef is constant. Then f is three-valued almost optimal if and only if f is almost optimal if and only if D_af is balanced for any $a \notin \{0, e\}$.

Thus the general form of codewords which correspond to an almost optimal Boolean function f which has a linear structure e is for odd m:

$$\Omega_f = (X^e + 1)\,\mathrm{y}$$

where y is the associated codeword of some bent function of $m - 1$ variables.

References

[1] A. Canteaut, C. Carlet, P. Charpin, C. Fontaine, Propagation characteristics and correlation-immunity of highly nonlinear Boolean functions, In *Advances in Cryptology- EUROCRYPT 2000*, Lecture Notes in Computer Science 1807, pp. 507-522. Springer-Verlag, 2000.

[2] A. Canteaut, C. Carlet, P. Charpin, C. Fontaine, On cryptographic properties of the cosets of $R(1, m)$, *IEEE Trans. Inform. Theory*, to appear.

[3] "Handbook of Coding Theory", chapter 11 (P. Charpin) and chapter 16 (E.F. Assmus and J. Key). V. S. Pless, W. C. Huffman, Eds, R. A. Brualdi, assistant editor, Amsterdam, the Netherlands: Elsevier, 1998.

[4] W. Meier and O. Staffelbach. Nonlinearity criteria for cryptographic functions. In *Advances in Cryptology - EUROCRYPT'89*, number 434 in Lecture Notes in Computer Science, pages 549–562. Springer-Verlag, 1990.

A Low Complexity Algorithm for the Construction of Algebraic Geometric Codes better than the Gilbert-Varshamov Bound*

K. Shum, I. Aleshnikov, P. V. Kumar and H. Stichtenoth

Abstract

Since the proof in 1982 by Tsfasman-Vladut and Zink, of the existence of algebraic geometric (AG) codes with asymptotic performance exceeding the Gilbert-Varshamov (G-V) bound, one of the challenges in coding theory has been to provide explicit descriptions for these codes. In a major step forward in 1995-96, Garcia and Stichtenoth provided an explicit description of algebraic curves such that AG codes constructed on them would have performance better than the G-V bound. However, the problem of constructing codes on the curves of Garcia and Stichtenoth has remained open. We present here for the first time, a low complexity algorithm for obtaining the generator matrix for AG codes on the curves of G-S. The symbol alphabet of the code is the finite field of $q^2, q^2 > 49$ elements. The complexity of the algorithm as measured in terms of multiplications and divisions over the finite field $GF(q^2)$ is *upper bounded* by $[N \log_q(N)]^3$ where N is the length of the code.

It is also shown that by concatenating the AG codes with short length binary block codes, one obtains long binary codes with very good asymptotic performance.

Address

K. Shum and P. V. Kumar,
Communication Sciences Institute, Electrical Engineering - Systems,
3740 McClintock Avenue EEB 534,
University of Southern California, Los Angeles, CA 90089-2565, USA
Phone: (213) 740 4668
Email: `keshum@milly.usc.edu`, `vijayk@ceng.usc.edu`

I. Aleshnikov and H. Stichtenoth
Fachbereich 6 Mathematik und Informatik,
Universität GH Essen,
D-45117 Essen, Germany.
Email: `mat314@uni-essen.de`, `stichtenoth@uni-essen.de`

---

*This research is supported by NSF under grant CCR-0073555

I Introduction

Long codes are judged on the basis of their parameters (δ, R) where δ is the fractional minimum distance and R is the code rate i.e., if $[n, k, d_{\text{min}}]$ are the length, dimension and minimum distance of the code respectively, then

$$\delta = \frac{d_{\text{min}}}{n} \text{ and } R = \frac{k}{n}.$$

The best long codes lie in the region defined by the Gilbert-Varshamov lower bound and the McEliece, Rodemich, Rumsey and Welch upper bound. One of the challenges in coding theory has been the construction of codes with symbol alphabet size fixed at r and growing length whose performance exceeds that of the G-V bound. It is desirable to keep r small as this allows for simpler encoding and decoding. While it is known that there exist long binary ($r = 2$) alternant and concatenated codes that meet the G-V bound, no explicit description of these codes exists. The existence or otherwise of long binary codes with performance improving upon the G-V bound remains open. A similar statement was true in the non-binary ($r > 2$) case until the early 1980's. Around 1980, V. D. Goppa [8] used the theory of algebraic curves to construct a new family of codes, now referred to as algebraic geometric (AG) codes.

The length of an AG code defined over a curve of genus g, is roughly equal to the number N of *rational* points on the curve, i.e, equal to the number of points having coordinates in the finite field $\mathbb{F}_r$ of r elements over which the curve is defined. The performance of an AG code of length N is governed by the equation

$$R \geq 1 - \delta - \frac{g}{N},$$

and thus depends upon the ratio g/N. Good codes result when the ratio g/N is small. However, the Drinfeld-Vlăduţ (D-V) bound states that this ratio cannot be too small:

$$\liminf_{g,N\to\infty} \frac{g}{N} \geq \frac{1}{\sqrt{r}-1}.$$

In 1982, Tsfasman, Vlăduţ and Zink (T-V-Z) [13] showed the existence of curves, known as modular curves, whose g/N ratio achieves the D-V bound. For $r \geq 49$, the resulting AG codes had performance exceeding that of the Gilbert-Varshamov bound, a result that caused considerable excitement in the coding community. However, the modular curves in [13] did not have an explicit description and no simple algorithm for generating codes on these curves is known although a polynomial-time algorithm is described in [9]. In a major step forward in 1995-96, Garcia and Stichtenoth (G-S) [5, 6] building on ideas of Feng and Pellikaan showed that two explicitly described curves also achieve the D-V bound. Since then, other examples of asymptotically optimal towers have surfaced in a paper by Elkies [3] and in a third paper of Garcia and Stichtenoth [7].

The two curves described by Garcia and Stichtenoth in [5, 6] are defined over $\mathbb{F}_r$, r a perfect square, i.e., $r = q^2$ and take on the form in both cases, of a sequence of curves of increasing genus g and number of rational points N_g satisfying

$$\lim_{g\to\infty} \frac{g}{N_g} = \frac{1}{q-1}.$$

In [14], Voss and Høholdt provide generator matrices for codes constructed on the first three curves in the first sequence of curves of Garcia and Stichtenoth. In [4], Haché extends this result to the fourth curve in the same sequence for the particular case $q^2 = 16$.

Our interest here is on the second sequence of curves and the kth curve in this sequence is defined by the equations

$$x_l^q + x_l = \frac{x_{l-1}^q}{x_{l-1}^{q-1} + 1}, \quad l = 1, 2, 3, \ldots k - 1 \ .$$

In the typical construction of a code on an algebraic curve, a particular rational point Q on an algebraic curve is singled out. The collection of all functions defined on the curve, whose only pole is Q with pole order upper bounded by a fixed integer m, forms a vector space $\mathcal{L}(mQ)$. If $\{f_1, f_2, \ldots, f_k\}$ is a basis over $\mathbb{F}_{q^2}$ for this vector space, then the ith, $1 \leq i \leq k$ row of the $(k \times N)$ generator matrix G_m of the AG code constructed on the curve is obtained by evaluating the function f_i at all rational points on the curve other than Q. Thus the integer m is a parameter of code construction, one that governs the dimension and minimum distance of the AG code.

So, given a curve and a rational point Q on the curve, one needs to find a basis for the vector spaces $\{\mathcal{L}(mQ) \mid m \geq 0\}$ to be able to build codes of varying dimension and minimum distance. The problem of finding a basis for the spaces $\{\mathcal{L}(mQ)\}$ in the case of the curves exhibited by Garcia and Stichtenoth has proved to be a challenging one. In [10], the authors determine for every integer m, the dimension of the vector space $\mathcal{L}(mQ)$. In [1], Aleshnikov, Deolalikar, Kumar and Stichtenoth identify a simply-described set of functions $\mathcal{A}$ whose span includes the vector spaces $\mathcal{L}(mQ)$. By this we mean that any function in any of the vector spaces $\mathcal{L}(mQ)$ can be expressed as a linear combination of the elements of $\mathcal{A}$.

In the present paper, we first identify a different and significantly smaller set $\mathcal{B}$ of functions with the property that their span includes the vector spaces $\mathcal{L}(mQ)$. The smaller size of this set allow us to provide a low complexity algorithm for constructing the generator matrix of the code. The algorithm proceeds as follows. The functions in $\mathcal{B}$ happen to have undesired poles at a collection B of points on the curve distinct from Q. A power series expansion is determined for each function at each of the points in B. Simple Gaussian elimination is then used to carry out pole cancellation, i.e., is used to determine the precise linear combinations of elements in $\mathcal{B}$ whose poles are confined to the single rational point Q. The functions so obtained can be shown to form a basis for the vector spaces $\mathcal{L}(mQ)$. The complexity of this algorithm, measured in terms of multiplications and divisions over the finite field $\mathbb{F}_{q^2}$ is upper bounded by $[N \log_q(N)]^3$ where N is the code length.

The nonbinary codes with performance exceeding the G-V bound can be concatenated with short binary block codes to yield long binary codes that also have good asymptotic performance. Examples of such binary codes obtained through concatenation when the outer AG code has symbol alphabet size $q^2 = 2^b$ not exceeding 256 are also presented (see Fig. 1). In the figure, the convex hull of the straight lines represents the best performance offered by these concatenated codes and as the figure shows, this is not far from the G-V bound.

II Background

We introduce the G-S curves in the language of function fields. From this viewpoint, the G-S curve is a tower $\{F_0, F_1, \cdots, F_n\}$ of Artin-Schreier extensions of the rational function

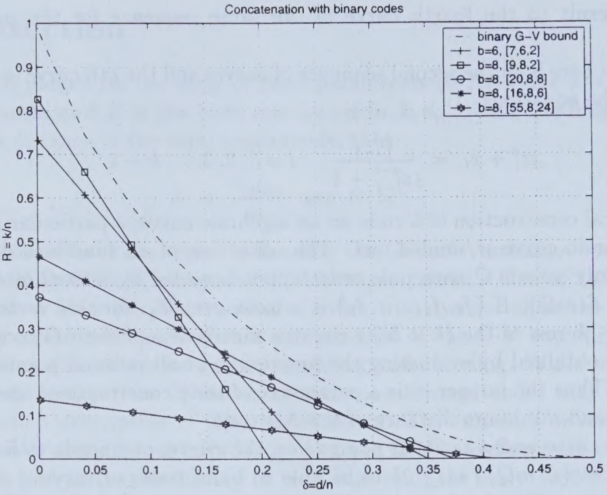


Figure 1: Asymptotic performance of concatenated codes

field $K(x_0)$, $K = \mathbb{F}_{q^2}$ given by

$$F_0 = K(x_0) \tag{1}$$

$$F_k = F_{k-1}(x_k) \quad \text{where } x_k^q + x_k = \frac{x_{k-1}^q}{x_{k-1}^{q-1}+1}, \quad 1 \le k \le n. \tag{2}$$

Let $P_\theta^{(0)}$, $\theta \in \mathbb{F}_{q^2}$, denote the unique zero of $x_0 - \theta$ in F_0. Let $P_\infty^{(0)}$ denote the unique pole of x_0 in $K(x_0)$. The place $P_\infty^{(0)}$ is completely ramified in the function field tower. Let $P_\infty^{(n)}$ be the unique place in F_n lying above $P_\infty^{(0)}$. The behavior in the tower of the places lying above $P_0^{(0)}$ is more complicated (see Fig. 2).

Some known properties of the G-S tower, taken mostly from [6] are listed below.

Field of Constants $\mathbb{F}_{q^2}$ is the full constant field for all F_k, $0 \le k \le n$.

Ramification Let

$$\Omega = \{\alpha \in \mathbb{F}_{q^2} \,|\, \alpha^q + \alpha = 0\}$$
$$\Omega^* = \Omega \setminus \{0\}.$$

Ramification in the tower takes place only above the places $P_\infty^{(0)}$ and $P_\alpha^{(0)}, \alpha \in \Omega$. The places $P_\infty^{(0)}$ and $P_\alpha^{(0)}$ for $\alpha \in \Omega^*$ are totally ramified throughout the tower (see Fig. 2). Ramification behavior above the place $P_0^{(0)}$ is discussed in the subsection below. For every $0 \le t \le n-1$, the places in F_t are either unramified in F_{t+1}/F_t or else, are totally (and wildly) ramified.

Genus The genus $g(F_n)$ of F_n is given by,

$$g(F_n) = \begin{cases} (q^{(n+1)/2} - 1)^2 & \text{for } n \text{ odd} \\ (q^{n/2} - 1)(q^{(n+2)/2} - 1) & \text{for } n \text{ even.} \end{cases}$$

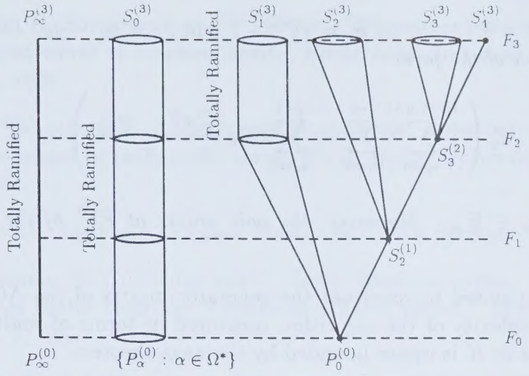

$P_\infty^{(3)}$
$S_0^{(3)}$
$S_1^{(3)}$
$S_2^{(3)}$
$S_3^{(3)}$
$S_4^{(3)}$
F_3
Totally Ramified
Totally Ramified
Totally Ramified
F_2
$S_3^{(2)}$
F_1
$S_2^{(1)}$
F_0
$P_\infty^{(0)}$
$\{P_\alpha^{(0)} : \alpha \in \Omega^*\}$
$P_0^{(0)}$

Figure 2: Ramification diagram for the tower.

Thus $g(F_n) \approx q^{n+1}$.

Number of places of degree one The places $P_\theta^{(0)}$, $\theta \in \mathbb{F}_{q^2} \setminus \Omega$ split completely in F_n/F_0. Thus the number $N(F_n)$ of places of degree 1 satisfies $N(F_n) \geq q^n(q^2 - q)$. Since

$$\lim_{n \to \infty} \frac{N(F_n)}{g(F_n)} = q - 1$$

the tower asymptotically meets the Drinfeld-Vlăduţ bound. The exact number of places of degree 1 for $n \geq 2$ is given by (see [2]),

$$N(F_n) = \begin{cases} q^n(q^2 - q) + 2q & \text{for odd } q, \\ q^n(q^2 - q) + 2q^2 & \text{for even } q. \end{cases}$$

Rings of Functions Define

$$R_n = \text{Integral closure of } \mathbb{F}_{q^2}[x_0] \text{ in } F_n.$$

Then R_n is the ring of functions that have pole only at $P_\infty^{(n)}$ and has the alternate description

$$R_n := \bigcup_{i \geq 0} \mathcal{L}(iP_\infty^{(n)}).$$

We will refer to an element of R_n as a *regular function.*

III The Main Theorems

The following result plays an important role in providing the low complexity algorithm:

Theorem 1. *For* $k = 0, \ldots, n$, *let* $g_k = x_k^{q-1} + 1$ *and* $\pi_k = g_0 \cdots g_k$. *Every function in* R_n *has an expression of the form*

$$x_0^l \left(\sum_{e_0=0}^{(n-1)q+1} \sum_{e_1=0}^{q-1} \cdots \sum_{e_n=0}^{q-1} c_{\mathbf{e}} g_0 \frac{x_0^{e_0} x_1^{e_1} \cdots x_n^{e_n}}{\pi_1 \cdots \pi_{n-1}} \right) \tag{3}$$

where $l \geq 0$ *and* $c_{\mathbf{e}} \in \mathbb{F}_{q^2}$. *Moreover, the pole orders at* $P_\infty^{(n)}$ *of the summands are pairwise distinct.*

An algorithm is devised to construct the generator matrix of the AG codes defined on F_n/K. The complexity of the algorithm measured in terms of multiplications and divisions of elements in K is upper bounded by the next theorem.

Theorem 2. *The generator matrix of the AG codes with length* $N = q^n(q^2-q)$ *associated to* $F_n/\mathbb{F}_{q^2}$ *can be computed with* $n^3N^3/(q-1)^2$ *operations in the field* K. *If* $n^3 < (q-1)^2$, *the complexity is upper bounded by* N^3.

References

[1] I. Aleshnikov, V. Deolalikar, P. V. Kumar and H. Stichtenoth, "Towards a basis for the space of regular functions in a tower of function fields meeting the Drinfeld-Vlăduţ bound", to appear in the proceedings of the *Fifth International Conference on Finite Fields and Applications*, August 2–6, 1999, University of Augsburg, Germany.

[2] I. Aleshnikov, P. V. Kumar, K. Shum and H. Stichtenoth, "On the Splitting of Places in a Tower of Function Fields Meeting the Drinfeld-Vlăduţ Bound", submitted to *IEEE Trans. Info. Theory.*

[3] N. Elkies, "Explicit modular towers," *Proc. 35th Annual Allerton Conference on Commun., Control and Computing*, Urbana, IL, 1997.

[4] G. Haché, *Construction effective des codes géométriques*, Thèse, Université Pierre & Marie Curie Paris VII, 1996.

[5] A. Garcia and H. Stichtenoth, "A tower of Artin-Schreier extensions of function fields attaining the Drinfeld-Vladut bound," *Invent. Math.*, vol. 121, 1995, pp. 211–222.

[6] A. Garcia and H. Stichtenoth, "On the Asymptotic Behaviour of Some Towers of Function Fields over Finite Fields," Journal of Number Theory, vol. 61, No. 2, December 1996, pp. 248–273.

[7] A. Garcia and H. Stichtenoth, "On Tame Towers over Finite Fields," preprint, 2000.

[8] V. D. Goppa, "Codes on algebraic curves," *Soviet Math. Dokl.*, vol. 24, No. 1, pp. 170–172, 1981.

[9] G. L. Katsman, M. Tsfasman and S. G. Vladut, "Modular curves and codes with a polynomial construction," *IEEE Trans. Inform. Theory*, vol. 30, (1984), pp. 353–355.

[10] R. Pellikaan, H. Stichtenoth and F. Torres, "Weierstrass semi-groups in an asymptotically good tower of function fields," Finite fields and their applications, vol 4, pp381–392, 1998.

[11] K. Shum, "Construction of long algebraic geometric codes better than the Gilbert-Varshamov bound in polynomial time", Ph.D. thesis, University of Southern California, 2000.

[12] H. Stichtenoth, *Algebraic Function Fields and Codes*, Springer-Verlag, 1993.

[13] M. A. Tsfasman, S. G. Vladut and T. Zink, "Modular curves, Shimura curves and Goppa codes better than the Varshamov-Gilbert bound," *Math. Nachr.*, vol. 109, (1982), pp. 21–28.

[14] C. Voss and T. Høholdt, "An explicit construction of a sequence of codes attaining the Tsfasman-Vladut-Zink bound. The first steps," *IEEE Trans. Inform. Theory*, vol 43, pp128–135, 1997.

Understanding Linear Codes of Constant Weight Using Virtual Linear Codes

Jay A. Wood
Department of Mathematics and Statistics
Western Michigan University
1903 W. Michigan Ave.
Kalamazoo, MI 49008–5248
`jay.wood@wmich.edu`
`http://unix.cc.wmich.edu/jwood/`

Abstract. In this paper we determine completely the structure of linear codes over $\mathbb{Z}/N\mathbb{Z}$ of constant weight. Namely, we determine exactly which modules underlie linear codes of constant weight, and we describe the coordinate functionals involved. The weight functions considered are: Hamming weight, Lee weight, and Euclidean weight. We prove a general uniqueness theorem for virtual linear codes of constant weight. Existence is settled on a case by case basis.

1. Introduction

This paper classifies the structure of linear codes of constant weight over $\mathbb{Z}/(N) = \mathbb{Z}/N\mathbb{Z}$. A linear code having constant weight means that every nonzero codeword has the same weight. Hamming, Lee, and Euclidean weights are all examined.

The classification specifies which modules over $\mathbb{Z}/(N)$ underlie linear codes of constant weight, and it specifies what the coordinate functionals need to be (up to an appropriate notion of equivalence).

There are a few surprises. While constant Hamming weight codes exist in all dimensions over finite fields, they almost never exist over the $\mathbb{Z}/(N)$'s that are not fields. Constant Lee or Euclidean weight codes exist for any module over $\mathbb{Z}/(2^\beta)$, but are comparatively rare over the $\mathbb{Z}/(N)$'s that have odd prime factors in N.

The structure of the proof is simple. There is the question of existence (given a module M over $\mathbb{Z}/(N)$, does it underlie a linear code of constant weight?), and there is the question of uniqueness (in how many ways can this occur, up to equivalence?).

We first prove a strong uniqueness theorem, Corollary 5.2, in the context of what we call virtual codes. Viewing linear codes from the linear functional viewpoint of [**1**], a linear code is a pair (M, η), where M is a module over R and $\eta : M^\sharp \to \mathbb{N}$ is a multiplicity function. Here, $M^\sharp = \mathrm{Hom}_R(M, R)$ is the linear dual of M, and η is keeping track of repeated columns in a generator martrix. Virtual codes, akin to virtual representations in representation theory, are pairs (M, η), where now $\eta : M^\sharp \to \mathbb{Q}$. The strong uniqueness theorem says that, for any M, there is a one-dimensional rational solution space of η's that correspond to constant weight codes. Only those η with values in $\mathbb{N}$ correspond to linear codes in the classical sense.

Armed with the uniqueness theorem, here is how the proof proceeds. Existence: guess what η should be and verify that it yields constant weight. By the uniqueness theorem,

η must be a basis for the solution space. If η has both positive and negative values, there is no classical solution. If η has all nonnegative values, clearing denominators leads to a classical solution η'.

Uniqueness: we are interested in the integral points in the one-dimensional rational solution space. The classical solution η' has nonnegative integral values. By dividing by the greatest common divisor of those values, we obtain a minimal integral solution η_0. Every other integral point in the solution space must be an integral multiple of η_0. Thus we see that constant weight codes, if they exist at all, must be replications of a minimal length model.

The proof of the uniqueness theorem depends on the extension theorem for weight preserving homomorphisms ([**7**], for example). Since the extension theorem is not known in general for many of the examples covered in this paper, some results include an extension property as part of their hypotheses. The existence results given do not depend on the extension theorem.

Here is a short guide to the contents of this paper. In Section 2 we discuss our ground rings R and modules over them. We also describe our (M, η) definition of linear codes. In Section 3 we introduce weight functions, the extension property, and viewing codes in terms of function spaces. Examples are then introduced.

In Sections 4 and 5 we introduce virtual codes and prove the strong uniqueness theorem. Sections 6–9 are concerned with existence: the basic strategy of the verifications, followed by precise statements of existence for Hamming, Lee, and Euclidean weights. Detailed verifications are omitted for lack of space.

2. Linear Codes over Finite Rings

Notational Conventions. In this paper, any ring denoted by R will be assumed to be a finite commutative ring with 1. The ideal generated by $r \in R$ will be denoted by (r). Any R-module M will be assumed to be finitely generated and unital, i.e., $1 \in R$ acts as the identity. The linear dual of M is denoted $M^\sharp := \mathrm{Hom}_R(M, R)$. The elements of $M^\sharp$ are the *linear functionals* on M. Integer residue rings are denoted $\mathbb{Z}/(N) := \mathbb{Z}/N\mathbb{Z}$. The natural numbers $\mathbb{N}$ will contain 0. The number of elements in a finite set S is $|S|$.

Let $R = \mathbb{Z}/(N)$, with prime factorization

$$N = p_1^{\beta_1} \dots p_l^{\beta_l}. \tag{2.1}$$

Every module M over R has a direct sum decomposition

$$M \cong \bigoplus_{i=1}^{l} \bigoplus_{j=1}^{\beta_i} \left(\mathbb{Z}/(p_i^j)\right)^{k_{i,j}}, \tag{2.2}$$

for appropriate non-negative integers $k_{i,j}$. Observe that

$$|M| = \prod_{i=1}^{l} p_i^{\sum_{j=1}^{\beta_i} j k_{i,j}}.$$

To avoid the situation where M is actually a pullback of a module defined over a quotient ring of R, we assume that, for all i,

$$k_{i,\beta_i} \geq 1. \tag{2.3}$$

There will be occasions where the prime number 2 will occur explicitly in (2.1). In that case $2 = p_0$ with exponent β_0 and integers $k_{0,j}$ in (2.2).

Linear codes will be described from the linear functional point of view of [**1**], although phrased in a slightly different way. A *linear code* C over R is a pair (M, η), where M is an R-module, the module *underlying* the code, and $\eta : M^\sharp \to \mathbb{N}$ is a *multiplicity function.* The *length* n of the linear code C is $n = \sum_{\lambda \in M^\sharp} \eta(\lambda)$. A linear code is *nondegenerate* if the multiplicity of the zero functional vanishes, i.e., if $\eta(0) = 0$.

A linear code (M, η) determines a linear homomorphism $\phi_\eta : M \to R^n$, $x \mapsto (\lambda(x))_{\lambda \in M^\sharp}$, where the entry $\lambda(x)$ appears $\eta(\lambda)$ times. The image of ϕ_η is a submodule of R^n, and this submodule is a linear code in the classical sense. Note that in defining ϕ_η one must choose an order in which to write down the terms $\lambda(x)$.

We shall usually assume that (M, η) satisfies the *coding axiom,* which states that ϕ_η is injective. By passing to a quotient, $M/\ker \phi_\eta$, the coding axiom holds automatically.

One way to view the definition above is to recall that a linear code is determined by its generator matrix G. The columns of G are given by linear functionals. Up to permutations of coordinate positions (the choice of order in writing down the terms in ϕ_η), the code is determined by the multiplicities of the various columns of G. It is exactly this information which is encoded by the multiplicity function η.

In an analogy with representation theory, we will have occasion to consider *virtual* linear codes. These are pairs (M, η), as above, where we allow η to have values in $\mathbb{Z}$ or $\mathbb{Q}$. That is, we allow linear functionals to occur with negative or rational multiplicities. More details will appear in Section 4.

3. Weight Functions and the Extension Property

In order to define the weight of codewords, we first define a *weight function* w on the ring R by assigning real number weights a_r to every $r \in R$. We assume that $a_0 = 0$ and that $a_r > 0$ for $r \neq 0$. This choice of weight function on R allows us to define a *weight function* $w_\eta : M \to \mathbb{R}$ on any linear code $C = (M, \eta)$:

$$w_\eta(x) = \sum_{\lambda \in M^\sharp} \eta(\lambda)\, a_{\lambda(x)}, \quad x \in M. \tag{3.1}$$

For example, Hamming weight uses $a_r = 1$, for all $r \neq 0$. We say that a linear code C has *constant weight* $L > 0$ if $w_\eta(x) = L$ for all nonzero $x \in M$. Since the zero element $0 \in M$ always has $w_\eta(0) = 0$, we hope the reader will tolerate this slightly misleading terminology.

To capture some of the symmetry of a weight function w, define the *symmetry group* of w to be

$$\mathrm{Sym}(w) := \{u \in \mathcal{U}(R) : a_{ur} = a_r, \text{ all } r \in R\},$$

where $\mathcal{U}(R)$ is the group of units of R. The group $\mathrm{Sym}(w)$ acts on both M and $M^\sharp$ by scalar multiplication, thereby decomposing M and $M^\sharp$ into $\mathrm{Sym}(w)$-orbits. Denote the $\mathrm{Sym}(w)$-orbits of $x \in M$ and $\lambda \in M^\sharp$ by $\mathrm{orb}(x)$ and $\mathrm{orb}(\lambda)$, respectively.

If $f : M' \to M$ is a morphism of R-modules, then the equation $u(\lambda \circ f) = (u\lambda) \circ f$ shows that the induced morphism $f^\sharp : M^\sharp \to M'^\sharp$ takes $\mathrm{Sym}(w)$-orbits on $M^\sharp$ to $\mathrm{Sym}(w)$-orbits on $M'^\sharp$.

LEMMA 3.1. *Suppose (M, η) is a linear code over R. If $x, y \in M$ satisfy $y \in \mathrm{orb}(x)$, then $w_\eta(y) = w_\eta(x)$.*

PROOF. If $y = ux$ for some $u \in \mathrm{Sym}(w)$, then $a_{\lambda(y)} = a_{u\lambda(x)} = a_{\lambda(x)}$, by the definition of $\mathrm{Sym}(w)$. The result follows immediately. ☐

Consider a linear code (M, η) over R. For any $\lambda \in M^\sharp$, define

$$\eta_S(\lambda) := \sum_{\mu \in \mathrm{orb}(\lambda)} \eta(\mu). \tag{3.2}$$

Then $\eta_S(\lambda)$ is the total multiplicity of linear functionals which belong to $\operatorname{orb}(\lambda)$. Clearly, $\eta_S(\lambda) = \eta_S(\mu)$ if $\mu \in \operatorname{orb}(\lambda)$. Note that (3.1) can be rewritten as

$$w_\eta(x) = \sum_{\lambda:\mathrm{rep}} \eta_S(\lambda)\, a_{\lambda(x)}, \quad x \in M, \tag{3.3}$$

where the summation is over one representative λ of each $\operatorname{Sym}(w)$-orbit. As we remarked above, the terms on the right side of (3.3) are independent of the choice of representatives for the $\operatorname{Sym}(w)$-orbits.

Two linear codes $C' = (M, \eta')$, $C = (M, \eta)$, are *scale equivalent* if $\eta'_S = \eta_S$. Let $\mathcal{O}$, $\mathcal{O}^\sharp$ denote the sets of nonzero $\operatorname{Sym}(w)$-orbits on M, $M^\sharp$, respectively. Denote the set of all functions $\mathcal{O}^\sharp \to \mathbb{N}$ by $\mathbb{N}[\mathcal{O}^\sharp]$.

THEOREM 3.2. *Fix an R-module M. There is a bijection between the set of all nondegenerate linear codes (M, η), up to scale equivalence, and the function space $\mathbb{N}[\mathcal{O}^\sharp]$.*

PROOF. To every code (M, η) we associate the function $\eta_S \in \mathbb{N}[\mathcal{O}^\sharp]$. Scale equivalent codes give rise to the same function η_S.

Conversely, given a function $g \in \mathbb{N}[\mathcal{O}^\sharp]$, we define a code as follows. For every $\operatorname{Sym}(w)$-orbit on $M^\sharp$, choose a representative. Then define $\eta : M^\sharp \to \mathbb{N}$ by

$$\eta(\lambda) = \begin{cases} g(\operatorname{orb}(\lambda)), & \lambda \text{ is a chosen representative,} \\ 0, & \text{otherwise.} \end{cases}$$

Then (M, η) is a nondegenerate linear code with $\eta_S = g$. A different choice of representatives for the $\operatorname{Sym}(w)$-orbits results in a scale equivalent code. ☐

Two linear codes $C' = (M', \eta')$, $C = (M, \eta)$, are *equivalent* if there exists an isomorphism $f : M' \to M$ such that $(M, \eta' \circ f^\sharp)$ and (M, η) are scale equivalent. If C' and C are equivalent, a re-indexing argument shows that $w_{\eta'}(x) = w_\eta(f(x))$ for all $x \in M'$. That is, f induces a weight-preserving isomorphism between the codes. The converse of this statement is discussed next.

DEFINITION EP. A weight function w over a ring R has the *extension property* (EP) if:

For any two linear codes $C' = (M', \eta')$, $C = (M, \eta)$ over R with an isomorphism $f : M' \to M$ satisfying $w_{\eta'}(x) = w_\eta(f(x))$ for all $x \in M'$, it follows that $(M, \eta' \circ f^\sharp)$ and (M, η) are scale equivalent. (Thus C' and C are equivalent via f.)

REMARK 3.3. Classically, two linear codes in R^n are equivalent if there is a $\operatorname{Sym}(w)$-monomial transformation on R^n taking one code to the other. ($\operatorname{Sym}(w)$-monomial transformations are those whose units belong to $\operatorname{Sym}(w)$.) In the present definition, multiplication by units in $\operatorname{Sym}(w)$ is handled by scale equivalence. Permutations play a role only when one defines the map $\phi_\eta : M \to R^n$, because one must choose an order in which to write down the terms $\lambda(x)$. Classically, it is the image of ϕ_η that is the code, so that the particular parameterization is not relevant. The isomorphism f allows one to change parameterizations. This is analogous to changing basis within the code, as when one treats classical equivalence via generator matrices.

If EP is satisfied, it follows that every weight preserving automorphism of R^n is a $\operatorname{Sym}(w)$-monomial transformation.

THEOREM 3.4. *Suppose R, w satisfy EP and that M is a fixed R-module. Define a mapping $W : \mathbb{N}[\mathcal{O}^\sharp] \to \mathbb{R}[\mathcal{O}]$, $g \mapsto w_g$, where*

$$w_g(\operatorname{orb}(x)) = \sum_{\lambda:\mathrm{rep}} g(\operatorname{orb}(\lambda))\, a_{\lambda(x)},$$

and the summation is over one representative λ *of each* $\mathrm{Sym}(w)$*-orbit. Then* W *is injective.*

PROOF. First observe that the terms on the right side of the expression for w_g do not depend on the choice of representatives for the $\mathrm{Sym}(w)$-orbits, nor does the right side depend upon the choice of representative for $\mathrm{orb}(x)$. Cf., (3.3) and Lemma 3.1.

Take any $g, h \in \mathbb{N}[\mathcal{O}^\sharp]$ with $W(g) = W(h)$. Let (M, η_g), (M, η_h) be linear codes corresponding to g, h as in Theorem 3.2. Then $W(g) = w_g$ is just the weight function of the code (M, η_g), as in (3.3). The equality $W(g) = W(h)$ means that $f = \mathrm{id}_M$ is a weight preserving isomorphism from (M, η_g) to (M, η_h). By EP, these two codes are scale equivalent. But that implies $g = h$, as desired. $\square$

We conclude this section with various examples of weight functions, their symmetry groups, and information about whether EP holds.

EXAMPLE 3.5. Hamming weight. Over any ring R, set $a_r = 1$ for $r \neq 0$; $a_0 = 0$. The symmetry group $\mathrm{Sym}(w) = \mathcal{U}(R)$, the full group of units of R. EP holds over finite fields [**4**] and over any finite Frobenius ring [**5**].

EXAMPLE 3.6. Lee weight on $R = \mathbb{Z}/(N)$. Choose representatives in the range $-N/2 < r \leq N/2$, and set $a_r = |r|$. It follows that $\mathrm{Sym}(w) = \{\pm 1\}$. EP has been numerically verified for $N \leq 256$ (MAPLE computations of the author which verify the sufficient condition of [**7**, Theorem 3.1]). EP also holds for rings of the form $\mathbb{Z}/(2^\beta)$, $\mathbb{Z}/(3^\beta)$, and for finite fields $\mathbb{F}_p$ with $p = 2q + 1$, q prime (work in preparation). We conjecture that EP holds for all N.

EXAMPLE 3.7. Euclidean weight on $R = \mathbb{Z}/(N)$. Set $a_r = |r|^2$, where representatives lie in the range $-N/2 < r \leq N/2$. Same as for Lee weight, $\mathrm{Sym}(w) = \{\pm 1\}$. EP has the same status and conjecture as for Lee weight.

4. Virtual Codes

We saw in Theorem 3.2 that linear codes, up to scale equivalence, are described by functions $\mathcal{O}^\sharp \to \mathbb{N}$. The function space $\mathbb{N}[\mathcal{O}^\sharp]$ is a semiring which we will embed into a ring, following similar ideas in representation theory, K-theory, Grothendieck groups, etc. We could use $\mathbb{Z}[\mathcal{O}^\sharp]$, but we find it more convenient to use $\mathbb{Q}[\mathcal{O}^\sharp]$, so as to utilize the power of linear algebra over a field. These considerations motivate the following definition.

A *virtual linear code* over a ring R is a pair (M, η), where M is an R-module and $\eta : M^\sharp \to \mathbb{Q}$ is a multiplicity function. Linear codes where η takes values in $\mathbb{N}$ will be called *classical* linear codes. The *weight* of an element $x \in M$ is defined exactly as before; see (3.1).

The definitions of nondegenerate, scale equivalence, and equivalence carry over verbatim from their classical counterparts. For reference, we state the virtual version of Theorem 3.2.

THEOREM 4.1. *Suppose* M *is a fixed* R*-module. Then there is a bijection between the set of nondegenerate virtual linear codes* (M, η)*, up to scale equivalence, and the function space* $\mathbb{Q}[\mathcal{O}^\sharp]$.

5. A Uniqueness Theorem

In this section we use Theorem 3.4 to prove an isomorphism theorem for virtual linear codes with rational weights. In turn, we obtain a uniqueness theorem for virtual linear codes of constant weight.

THEOREM 5.1. *Assume R, w satisfy EP and that $a_r \in \mathbb{Q}$ in (3.1). Fix an R-module M. Then the mapping*

$$W : \mathbb{Q}[\mathcal{O}^\sharp] \to \mathbb{Q}[\mathcal{O}],$$

defined as in Theorem 3.4, is an isomorphism of $\mathbb{Q}$-vector spaces.

PROOF. The rationality of the a_r guarantees that $w_g \in \mathbb{Q}[\mathcal{O}] \subset \mathbb{R}[\mathcal{O}]$.

Since $\mathrm{Sym}(w)$ acts by scalar multiplication, the vector spaces $\mathbb{Q}[\mathcal{O}]$, $\mathbb{Q}[\mathcal{O}^\sharp]$ have the same finite dimension. Because W is linear, it suffices to show that W is injective.

Take any $g, h \in \mathbb{Q}[\mathcal{O}^\sharp]$ with $W(g) = W(h)$. By clearing denominators in all the values of g, h, we see that there is a positive integer B such that $Bg, Bh \in \mathbb{Z}[\mathcal{O}^\sharp]$. By adding a sufficiently large integer D to all the values of Bg, Bh, we obtain $Bg+D, Bh+D \in \mathbb{N}[\mathcal{O}^\sharp]$. Since $W(Bg + D) = W(Bh + D)$ follows from $W(g) = W(h)$, Theorem 3.4 now implies that $Bg + D = Bh + D$. Thus $g = h$. ☐

COROLLARY 5.2. *Assume the conditions of Theorem 5.1, and fix an R-module M. Then the set of scale equivalence classes of nondegenerate virtual linear codes (M, η) of constant weight over R forms a one-dimensional subspace of $\mathbb{Q}[\mathcal{O}^\sharp]$.*

PROOF. Nondegenerate constant weight codes correspond to the constant functions in $\mathbb{Q}[\mathcal{O}]$. ☐

THEOREM 5.3. *Suppose (M, η') and (M, η) both have constant weight and that they are equivalent via $f \in \mathrm{Aut}(M)$. Then they are scale equivalent. Thus, for linear codes of constant weight, scale equivalence classes are the same as equivalence classes.*

PROOF. Observe that $w_{\eta' \circ f^\sharp}(x) = w_{\eta'}(f^{-1}(x)) = w_{\eta'}(x)$, since (M, η') has constant weight. By hypothesis, $(M, \eta' \circ f^\sharp)$ is scale equivalent to (M, η), so that $w_{\eta' \circ f^\sharp} = w_\eta$. Thus $w_{\eta'} = w_\eta$, and (M, η'), (M, η) are scale equivalent. ☐

We now interpret Corollary 5.2 in classical terms. Given a linear code $C = (M, \eta)$, the d-fold *replication* of C is the linear code $dC = (M, d\eta)$, i.e., every multiplicity $\eta(\lambda)$ is multiplied by a factor of d. In classical coding terminology, we repeat d times each column of a generator matrix for C.

THEOREM 5.4. *Assume the conditions of Theorem 5.1. If M underlies a classical linear code of constant weight, then there is a nondegenerate classical linear code (M, η) of constant weight which has minimal length, and it is unique up to equivalence. Any other nondegenerate classical linear code (M, η') of constant weight is a d-fold replication of (M, η), up to equivalence.*

Moreover, if the multiplicity function η of a virtual linear code (M, η) of constant weight attains both positive and negative values, then there is no *classical linear code of constant weight with underlying module M.*

PROOF. Let H be the one-dimensional subspace of $\mathbb{Q}[\mathcal{O}^\sharp]$ consisting of virtual linear codes of constant weight. By hypothesis, $H \cap \mathbb{N}[\mathcal{O}^\sharp]$ contains a nonzero element μ. Dividing μ by the gcd of its values yields a nonzero $\eta \in H \cap \mathbb{N}[\mathcal{O}^\sharp]$ of minimal length. Any other element of $H \cap \mathbb{N}[\mathcal{O}^\sharp]$ is an integral multiple of η.

If η attains both positive and negative values, then $H \cap \mathbb{N}[\mathcal{O}^\sharp]$ is zero. ☐

6. Existence: Basic Strategy

Because of Corollary 5.2, questions of existence boil down to "guess and check": guess what the answer should be and then check that it is correct. All the guesses are based on extensive calculations—none of which appear in this paper. The verifications of the guesses are intricate, and they are omitted due to lack of space.

Underlying the verifications is a simple observation. Suppose $E \subset M^\sharp$ is a submodule of linear functionals and $x \in M$. Then $\check{x} : E \to R$, $\lambda \mapsto \lambda(x)$, is an R-linear homomorphism. Its image $\operatorname{im} \check{x} \subset R$ is an ideal, and every element of $\operatorname{im} \check{x}$ is hit $|\ker \check{x}|$ times. Thus,

$$\sum_{\lambda \in E} a_{\lambda(x)} = |\ker \check{x}| \sum_{r \in \operatorname{im} \check{x}} a_r = \frac{|E|}{|\operatorname{im} \check{x}|} \sum_{r \in \operatorname{im} \check{x}} a_r. \tag{6.1}$$

We then show that expressions built up from sums of this type are independent of the choice of nonzero x. The exact expressions for the sums $\sum_{r \in \operatorname{im}(\check{x})} a_r$ depend heavily upon the particular weight function used.

Assuming the validity of EP, Theorem 5.4 and its proof provide the classification of constant weight codes, once questions of existence have been settled. This information will not be repeated for every example.

7. Hamming Weight

THEOREM 7.1. *Let $R = \mathbb{Z}/(N)$, and let M be any module over R. We assume the notation in (2.1) and (2.2), in particular that $k_{i,\beta_i} \geq 1$. Set $K_i := \sum_{j=1}^{\beta_i} k_{i,j}$. For every nonzero $\lambda \in M^\sharp$, assign the multiplicity*

$$\eta(\lambda) = \prod_{\substack{i: \\ \lambda \in p_i M^\sharp}} \left(1 - p_i^{K_i - 1}\right).$$

The resulting virtual linear code has constant Hamming weight

$$|M| \prod_{i=1}^{l} (1 - 1/p_i).$$

COROLLARY 7.2. *Over $R = \mathbb{Z}/(N)$, an R-module M as in (2.2) underlies a classical linear code of constant Hamming weight only in the following circumstances:*

- *$N = p$, a prime, in which case $\eta(\lambda) = 1$ for all nonzero $\lambda \in M^\sharp$, see Remark 7.3, or*
- *M is free of rank 1.*

REMARK 7.3. Let us examine carefully the case where R is the finite field $\mathbb{F}_q$. Suppose M is a k-dimensional vector space over $\mathbb{F}_q$. Since $\operatorname{Sym}(w) = \mathbb{F}_q^\times$, we see that every nonzero $\operatorname{Sym}(w)$-orbit has $q - 1$ elements. Thus $\eta_S(\lambda) = q - 1$ for all nonzero λ. The proof of Theorem 5.4 shows that the shortest length code of this dimension has $\eta_S(\lambda) = 1$.

In classical terms, the code has a generator matrix whose columns consist of one representative from each one-dimensional subspace of $\mathbb{F}_q^k$. This reproves a classical result [2].

8. Lee Weight

We continue with the notation used in Section 7.

THEOREM 8.1. *Let $R = \mathbb{Z}/(N)$ and M be as above. To every nonzero $\lambda \in M^\sharp$, assign the multiplicity*

$$\eta(\lambda) = \prod_{\substack{p_i \text{ odd:} \\ \lambda \in p_i M^\sharp}} \left(1 - p_i^{K_i - 2}\right).$$

The resulting virtual linear code has constant Lee weight

$$(N/4)|M| \prod_{i:\, p_i \text{ odd}} (1 - 1/p_i^2).$$

COROLLARY 8.2. *Over $R = \mathbb{Z}/(N)$, an R-module M underlies a classical linear code of constant Lee weight only in the following circumstances:*

- *$N = p$, a prime, M arbitrary, in which case $\eta(\lambda) = 1$ for all nonzero $\lambda \in M^\sharp$.*
- *$N = 2^{\beta_0}$, M arbitrary, $\eta(\lambda) = 1$: Carlet,* [**3**].
- *N arbitrary, but M restricted by $K_i \le 2$, for all i with p_i odd.*

9. Euclidean Weight

We continue to work over $R = \mathbb{Z}/(N)$, with prime factorization $N = 2^{\beta_0} p_1^{\beta_1} \cdots p_l^{\beta_l}$ (the p_i being odd primes).

For ease of exposition, we will consider several cases, starting with the case where N is odd (i.e., $\beta_0 = 0$).

LEMMA 9.1. *Suppose N is odd. When summing over ideals of R, Euclidean weight is proportional to Lee weight.*

PROOF. Let $r|N$, with $ur = N$. We calculate $\sum_{s \in (r)} a_s$, for both Lee and Euclidean weights. Exploiting $\pm$-symmetry, we find that

$$\begin{aligned} \sum_{s \in (r)} a_s &= 2 \sum_{t=1}^{(u-1)/2} a_{tr} \\ &= \begin{cases} 2(r + 2r + \cdots + r(u-1)/2) & \text{Lee} \\ 2(r^2 + (2r)^2 + \cdots + (r(u-1)/2)^2) & \text{Euclidean} \end{cases} \\ &= \begin{cases} r(u^2-1)/4 & \text{Lee} \\ (N/3) \cdot r(u^2-1)/4 & \text{Euclidean.} \quad \square \end{cases} \end{aligned}$$

THEOREM 9.2. *A virtual linear code over $\mathbb{Z}/(N)$, N odd, has constant Euclidean weight if and only if it has constant Lee weight. For N odd, constant Euclidean weight codes are given by Theorem 8.1 and Corollary 8.2. The weights are multiplied by a factor of $N/3$.*

We now turn to the cases where N is even ($\beta_0 > 0$). The answers depend on whether N is a power of 2 or not. In either case, we need to filter the module $M^\sharp$:

$$M^\sharp \supset 2M^\sharp \supset \cdots \supset 2^{\beta_0 - 1} M^\sharp \supset 2^{\beta_0} M^\sharp. \tag{9.1}$$

For nonzero $\lambda \in M^\sharp$, define $\nu(\lambda)$ to be the largest integer $\le \beta_0$ such that $\lambda \in 2^{\nu(\lambda)} M^\sharp$.

We define some new quantities in terms of the numbers $k_{0,j}$ of (2.3). Set $e_0 = 1$ and

$$e_i = k_{0,1} + 2k_{0,2} + \cdots + (i-1)k_{0,i-1} + i(k_{0,i} + \cdots + k_{0,\beta_0}) - i, \tag{9.2}$$

for $1 \le i \le \beta_0$.

Let us now turn to the case where $N = 2^{\beta_0}$ is a power of 2. In this situation, $2^{\beta_0} M^\sharp = 0$ in (9.1), so that $\nu(\lambda) \le \beta_0 - 1$, for all nonzero $\lambda \in M^\sharp$.

THEOREM 9.3 ([**6**]). *Let $R = \mathbb{Z}/(N)$, $N = 2^{\beta_0}$; let M be a module over R, as above. For every nonzero $\lambda \in M^\sharp$, assign the multiplicity*

$$\eta(\lambda) = \sum_{i=0}^{\nu(\lambda)} 2^{e_i}.$$

The resulting linear code is classical and has constant Euclidean weight $2^{2\beta_0-2}|M| = (N^2/4)|M|$ *and length*

$$n = \frac{|M|}{2^{\beta_0-1}}\left(3\cdot 2^{\beta_0-1}-1\right) - \sum_{i=0}^{\beta_0-1} 2^{e_i}.$$

Finally, we consider the general even case where $N = 2^{\beta_0}p_1^{\beta_1}\cdots p_l^{\beta_l}$, $\beta_0 > 0$, $l \geq 1$. Using (9.2), define

$$e_0' = e_0, \ldots, e_{\beta_0-1}' = e_{\beta_0-1}, e_{\beta_0}' = e_{\beta_0}+1.$$

As above, set $K_i = \sum_{j=1}^{\beta_i} k_{i,j}$.

THEOREM 9.4. *Suppose* $R = \mathbb{Z}/(N)$, $N = 2^{\beta_0}p_1^{\beta_1}\cdots p_l^{\beta_l}$, $\beta_0 > 0$, $l \geq 1$, M *as above. For every nonzero* $\lambda \in M^\sharp$, *assign the multiplicity*

$$\eta(\lambda) = \left(\sum_{i=0}^{\nu(\lambda)} 2^{e_i'}\right) \prod_{\substack{p_i \text{ odd:} \\ \lambda \in p_i M^\sharp}} \left(1 - p_i^{K_i-2}\right).$$

The resulting virtual linear code has constant Euclidean weight

$$(N^2/4)|M| \prod_{i=1}^{l} (1 - 1/p_i^2).$$

COROLLARY 9.5. *Over* $R = \mathbb{Z}/(N)$, *an* R*-module* M *underlies a classical linear code of constant Euclidean weight only in the following circumstances:*

- $N = p$, *a prime,* M *arbitrary, in which case* $\eta(\lambda) = 1$ *for all nonzero* $\lambda \in M^\sharp$.
- $N = 2^{\beta_0}$, M *arbitrary: Theorem* 9.3.
- N *arbitrary, but* M *restricted by* $K_i \leq 2$, *for all* $i = 1, 2, \ldots, l$.

ACKNOWLEDGMENTS. The author thanks Purdue University Calumet, the University of Notre Dame, and Western Michigan University for sabbatical support. I thank J. Wolfmann, Ph. Langevin, and the research group GRIM at Université Toulon-Var for their hospitality during the spring of 2000. I also thank H. N. Ward for some key ideas and E. S. Moore for many things.

References

1. E. F. Assmus, Jr. and H. F. Mattson, *Error-correcting codes: An axiomatic approach*, Inform. and Control **6** (1963), 315–330.
2. A. Bonisoli, *Every equidistant linear code is a sequence of dual Hamming codes*, Ars Combinatoria **18** (1984), 181–186.
3. C. Carlet, *One-weight* $\mathbb{Z}_4$*-linear codes*, Coding Theory, Cryptography and Related Areas (J. Buchmann, T. Høholdt, H. Stichtenoth, and H. Tapia-Recillas, eds.), Springer, Berlin, 2000, pp. 57–72.
4. F. J. MacWilliams, *Error-correcting codes for multiple-level transmission*, Bell System Tech. J. **40** (1961), 281–308.
5. J. A. Wood, *Duality for modules over finite rings and applications to coding theory*, Amer. J. Math. **121** (1999), 555–575.
6. ———, *Linear codes over* $\mathbb{Z}/(2^k)$ *of constant Euclidean weight*, Proceedings of the Thirty-Seventh Annual Allerton Conference on Communication, Control, and Computing, University of Illinois, 1999, pp. 895–896.
7. ———, *Weight functions and the extension theorem for linear codes over finite rings*, Finite fields: Theory, Applications and Algorithms (R. C. Mullin and G. L. Mullen, eds.), Contemp. Math., vol. 225, Amer. Math. Soc., Providence, 1999, pp. 231–243.

High-Rate Linear Space-Time Codes

BABAK HASSIBI   BERTRAND M. HOCHWALD
Bell Laboratories, Lucent Technologies
600 Mountain Avenue, Murray Hill, NJ 07974
`{hassibi,hochwald}@bell-labs.com`

Abstract

Multiple-antenna systems that operate at high rates require simple yet effective space-time transmission schemes to handle the large traffic volume in real time. At rates of tens of bits/sec/Hz, V-BLAST, where every antenna transmits its own independent substream of data, has been shown to have good performance and simple encoding and decoding. Yet V-BLAST suffers from its inability to work with fewer receive antennas than transmit antennas. Furthermore, because V-BLAST transmits independent data streams on its antennas there is no built-in spatial coding to guard against deep fades from any given transmit antenna. On the other hand, there are many previously-proposed space-time codes that have good fading resistance and simple decoding, but these codes generally have poor performance at high data rates or with many antennas.

We propose a high-rate coding scheme that can handle any configuration of transmit and receive antennas and that subsumes both V-BLAST and many proposed space-time codes as special cases. The scheme transmits substreams of data in linear combinations over space and time. The codes are designed to optimize the mutual information between the transmitted and received signals. Because of their linear structure, the codes retain the decoding simplicity of V-BLAST, and because of their information-theoretic optimality, they possess many coding advantages. We give examples of the codes and show that their performance is generally superior to earlier proposed methods over a wide range of rates and SNR's. The full version of this paper [1] is available for download at `http://mars.bell-labs.com`

1 Introduction and Model

It is widely acknowledged that reliable fixed and mobile wireless transmission of video, data, and speech at high rates will be an important part of future telecommunications systems. One way to get high rates on a scattering-rich wireless channel is to use multiple transmit and/or receive antennas. Early uses of multiple transmit antennas in a scattering environment achieve reliability through "diversity". To keep the transmitter and receiver complexity low, linear processing is often preferred [2]. To achieve the high data rates promised in [3], however, new approaches for space-time transmission are needed. One such approach is V-BLAST (Vertical Bell Labs Layered Space-Time) [4] which breaks the original data stream into substreams that are transmitted on the individual antennas. The receiver decodes the substreams using a sequence of nulling and cancelling steps.

Since then there has been considerable work on a variety of space-time transmission schemes and performance measures [5] such as the space-time trellis codes of [6] and the space-time block codes of [7, 8] for the known channel and [9, 10, 11] for the unknown channel. At very high rates and with a large number of antennas, many of these space-time codes suffer from complexity or performance difficulties. Although V-BLAST can handle high data rates with reasonable complexity, the decoding scheme presented in [4] does not work with fewer receive than transmit antennas.

We present a space-time transmission scheme that has many of the coding and diversity advantages of previously designed codes, but also has the decoding simplicity of V-BLAST at high data rates. The codes work with arbitrary numbers of transmit and receive antennas.

The codes break the data stream into substreams that are dispersed in linear combinations over space and time. We refer to them simply as LD codes (linear dispersion codes). They:

1. Subsume, as special cases, both V-BLAST [4] and the block codes of [8]
2. Generally outperform both
3. Can be used for any number of transmit and receive antennas
4. Are very simple to encode
5. Can be decoded using a variety of simple linear-algebraic techniques such as:
   (a) Successive nulling and cancelling (V-BLAST [4], square-root V-BLAST [12])
   (b) Sphere decoding [13]
6. Are designed with the numbers of both the transmit *and* receive antennas in mind
7. Satisfy an information-theoretic optimality criterion
   - the codes are designed to maximize the mutual information between the transmit and receive signals

We briefly summarize the general structure of the LD codes. Suppose that there are M transmit antennas, N receive antennas, and an interval of T symbols available to us during which the propagation channel is constant and known to the receiver. The transmitted signal can then be written as a $T \times M$ matrix S that governs the transmission over the M antennas during the interval. We assume that the data sequence has been broken into Q substreams and that $s_1, \ldots, s_Q$ are the complex symbols chosen from an arbitrary, say r-PSK or r-QAM, constellation. We call a rate $R = (Q/T)\log_2 r$ linear dispersion code one for which S obeys

$$S = \sum_{q=1}^{Q} \left(\alpha_q A_q + j\beta_q B_q\right), \tag{1}$$

where the real scalars $\{\alpha_q, \beta_q\}$ are given by $s_q = \alpha_q + j\beta_q, \quad q = 1, \ldots Q$. The code is completely specified by the fixed $T \times M$ complex matrices $A_1, \ldots, A_Q$ and $B_1, \ldots, B_Q$. Each individual codeword is determined by the complex scalars $\{s_1, \ldots, s_Q\}$.

Many known codes can be identified as special cases. The 2 × 2 code of [7]

$$S = \begin{bmatrix} s_1 & s_2 \\ -s_2^* & s_1^* \end{bmatrix}$$

for example, can be obtained as a special case of (1) by taking $T = M = Q = 2$ and

$$A_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad A_2 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}, \quad B_1 = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad B_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

As another example, V-BLAST in [4] transmits individual substreams on each antenna. There is no spatial channel code structure and hence the blocksize T is irrelevant. We can therefore take $T = 1$ and hence $S = [s_1, s_2, \ldots, s_M]$. This is obtained in (1) by taking $Q = M$ and

$$A_q = B_q = [\underbrace{0, \ldots, 0}_{q-1}, 1, \underbrace{0, \ldots, 0}_{M-q}], \qquad q = 1, \ldots, Q.$$

The design of LD codes depends crucially on the choices of the parameters T, Q and the dispersion matrices $\{A_q, B_q\}$. To choose the $\{A_q, B_q\}$ we propose to optimize a nonlinear information-theoretic criterion: namely, the mutual information between the transmitted signals $\{\alpha_q, \beta_q\}$ and the received signal.

1.1 The Multiple-Antenna Model

Consider a narrow-band, flat-fading, multi-antenna communication system with M transmit and N receive antennas, where the channel is constant for at least T channel uses. Then the transmitted and received signals can be related by

$$X = \sqrt{\frac{\rho}{M}} SH + V, \tag{2}$$

where $X \in \mathcal{C}^{T\times N}$ is the received matrix, $S \in \mathcal{C}^{T\times M}$ is the transmitted matrix, and $V \in \mathcal{C}^{T\times N}$ is the additive noise matrix with independent $\mathcal{CN}(0,1)$ entries. In X, S, and V, time runs vertically and space runs horizontally. $H \in \mathcal{C}^{M\times N}$ is the channel matrix comprising independent $\mathcal{CN}(0,1)$ entries. The entries of the transmitted matrix S are assumed to have, on the average, unit variance entries, which implies the power constraint $\mathrm{E}\,\mathrm{tr}\, SS^* = TM$ and, along with the normalization $\sqrt{\frac{\rho}{M}}$, that ρ is the SNR at the receiver, independently of M.

H is assumed to be known to the receiver but not to the transmitter, in which case the channel capacity (often referred to as the *perfect-knowledge* capacity) is [3, 14]

$$C(\rho, M, N) = \mathrm{E}\,\log\det\left(I_N + \frac{\rho}{M} HH^*\right). \tag{3}$$

2 Information-Theoretic Analysis of Some Space-Time Codes

To motivate the LD codes let us begin by reviewing some existing space-time codes.

2.1 Mutual information attainable with orthogonal designs

An *orthogonal design* is introduced by Alamouti in [7] for $T = M = 2$ and has the structure

$$S = \begin{bmatrix} s_1 & s_2 \\ -s_2^* & s_1^* \end{bmatrix}. \tag{4}$$

The complex scalars s_1 and s_2 are drawn from a particular (r-PSK or r-QAM) constellation, but we may simply assume that they are random variables such that $\mathrm{E}\left(|s_1|^2 + |s_2|^2\right) = 2$. One can see from (4) that S has orthogonal columns and, moreover, that any linear combination of S's also has orthogonal columns. The orthogonal design has various interesting properties, one of which is the fact that ML estimation of s_1 and s_2 can be performed independently.

We may ask how much mutual information the orthogonal design structure (4) can attain? To answer this question we need to compute $C_{orth}(\rho, N)$, the mutual information between the transmitted and received signals, and compare it with $C(\rho, M, N)$, the capacity of the $M \times N$ multi-antenna system. This is done in [1], with the results given below:

$$\begin{cases} C_{orth}(\rho, N=1) &= C(\rho, M=2, N=1) \\ C_{orth}(\rho, N=2) &= C(2\rho, M=4, N=1) \end{cases} \tag{5}$$

The above equations imply that the orthogonal design (4) can achieve full channel capacity of the $M=2, N=1$ system, but not of the $M=2, N=2$ system. Thus, when $N=2$ we incur a loss with the structure (4), which from Figure 1 can be substantial at high SNR.

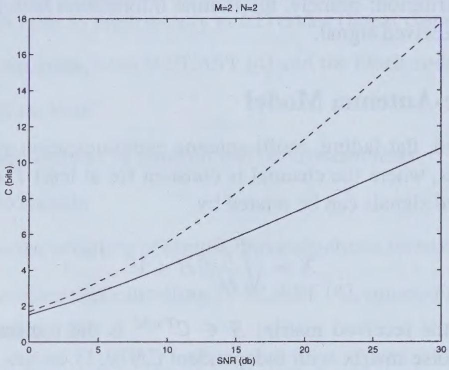


Figure 1: Solid line: maximum mutual information for 2 × 2 orthogonal design. Dashed line: capacity of $M=2$, $N=2$ system.

2.2 Mutual information attainable with V-BLAST

A scheme that is proven to be suitable for very high rates (tens of bits/sec/Hz) is V-BLAST [4], where each transmit antenna during each channel use sends an independent signal (often referred to as a substream). Thus, over a block of T channel uses,

$$S = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1M} \\ s_{21} & s_{22} & \cdots & s_{2M} \\ \vdots & \vdots & \ddots & \vdots \\ s_{T1} & s_{T2} & \cdots & s_{TM} \end{bmatrix}, \tag{6}$$

where each s_{ij} is an independent symbol, typically drawn from a PSK or QAM constellation.

When $N \geq M$ there exist efficient schemes for decoding the V-BLAST matrices. These are based on "successive nulling and cancelling" [4], and its more efficient variants [12], as well as more recently on sphere decoding [13]. The decoding methods have polynomial complexity in N, with the efficient implementations typically requiring $O(N^3)$ computations.

Because there is no restriction on the transmitted matrix S in (6), the maximum mutual information that can be achieved by transmitting V-BLAST-like matrices is indeed the full multi-antenna channel capacity. Nevertheless, V-BLAST suffers from two deficiencies. First, the simple decoding algorithm (nulling and cancelling) fails when there are fewer receive antennas than transmit antennas. Second, because V-BLAST transmits independent data streams on its antennas there is no built-in spatial or temporal coding.

3 Linear-Dispersion Space-Time Codes

We call a *linear-dispersion (LD) code* one for which

$$S = \sum_{q=1}^{Q} (\alpha_q A_q + j\beta_q B_q), \tag{7}$$

where $s_1, \ldots, s_Q$ are complex scalars (typically chosen from a r-PSK or r-QAM constellation) and where the A_q and B_q are *fixed* $T \times M$ complex matrices. Without loss of generality, we assume that $\alpha_1, \ldots, \alpha_Q$ and $\beta_1, \ldots, \beta_Q$ have variance $\frac{1}{2}$ and are uncorrelated. This induces the following normalization on the matrices $\{A_q, B_q\}$:

$$\sum_{q=1}^{Q} \left(\mathrm{tr}\, A_q^* A_q + \mathrm{tr}\, B_q^* B_q\right) = 2TM. \tag{8}$$

As mentioned earlier, the dispersion codes (7) subsume as special cases both orthogonal designs and V-BLAST. Note that in V-BLAST each signal $\{\alpha_q, \beta_q\}$ is transmitted from only one antenna during only one channel use. With the LD codes, however, the dispersion matrices potentially transmit some combination of *each* symbol from *each* antenna at *every* channel use. This can lead to desirable coding properties.

3.1 Decoding

An important property of the LD codes (7) is their linearity in the variables $\{\alpha_q, \beta_q\}$. This leads to the following linear relation between the $\{\alpha_q, \beta_q\}$ and the received signals:

$$\underbrace{\begin{bmatrix} x_{R,1} \\ x_{I,1} \\ \vdots \\ x_{R,N} \\ x_{I,N} \end{bmatrix}}_{\triangleq x} = \sqrt{\frac{\rho}{M}}\mathcal{H} \underbrace{\begin{bmatrix} \alpha_1 \\ \beta_1 \\ \vdots \\ \alpha_Q \\ \beta_Q \end{bmatrix}}_{\triangleq s} + \underbrace{\begin{bmatrix} v_{R,1} \\ v_{I,1} \\ \vdots \\ v_{R,N} \\ v_{I,N} \end{bmatrix}}_{\triangleq v}, \tag{9}$$

where $x_{R,n}$, $x_{I,n}$, $v_{R,n}$, and $v_{I,n}$, $n = 1, \ldots, N$, denote the real and imaginary parts of the columns of X and V, respectively, and where $\mathcal{H}$ is an *equivalent* $2NT \times 2Q$ real channel matrix whose entries depend on the entries of H and the dispersion matrices $\{A_q, B_q\}$—for an expression for $\mathcal{H}$ see [1].

Note that the equivalent channel $\mathcal{H}$ is known to the receiver because the original channel H, and the dispersion matrices $\{A_q, B_q\}$ are all known to the receiver. When $Q \leq NT$ there are more equations than unknowns and we may therefore use any decoding technique already in place for V-BLAST to decode the signals $\{\alpha_q, \beta_q\}$.

3.2 Design of the Dispersion Codes

Although we have introduced the LD structure we have not yet specified Q or the dispersion matrices $A_1, \ldots, A_Q$ and $B_1, \ldots, B_Q$. We have the inequality $Q \leq NT$. Intuitively, the larger Q is, the higher the maximum mutual information is since the matrix signal S has more degrees of freedom. On the other hand, the smaller Q is, the more of a coding effect we obtain. As a general practice, we find it useful to take $Q = \min(M, N) \cdot T$ since this tends to maximize the mutual information while still having the benefit of coding gain.

We are left with the question of how to design the dispersion matrices. Fortunately, information theory suggests a route: note that orthogonal designs are deficient in the maximum mutual information they support for $M > 2$ or $N > 1$. We therefore choose $\{A_q, B_q\}$ to maximize the mutual information between s and x in (9). This guarantees that we are taking the smallest possible mutual information penalty within the LD structure (7).

The Design Method

1. Choose $Q \leq NT$ (typically, $Q = \min(M, N) \cdot T$).

2. Choose $\{A_q, B_q\}$ that solve the optimization problem

$$C_{\mathrm{LD}}(\rho, T, M, N) = \max_{A_q, B_q, q=1,\ldots Q} \frac{1}{2T} \mathrm{E} \log\det\left(I_{2NT} + \frac{\rho}{M}\mathcal{H}\mathcal{H}^t\right) \tag{10}$$

   for an SNR ρ of interest, subject to one of the following constraints

   (i) $\sum_{q=1}^{Q}\left(\mathrm{tr}\, A_q^* A_q + \mathrm{tr}\, B_q^* B_q\right) = 2TM$

   (ii) $\mathrm{tr}\, A_q^* A_q = \mathrm{tr}\, B_q^* B_q = \frac{TM}{Q}, q = 1, \ldots Q$

   (iii) $A_q^* A_q = B_q^* B_q = \frac{T}{Q} I_M, q = 1, \ldots Q$

We now make some remarks:

1. Constraint (i) is simply the power constraint (8). Constraint (ii) is more restrictive and ensures that each of the transmitted signals α_q and β_q are transmitted with the same overall power. Constraint (iii) is the most restrictive, since it forces the symbols α_q and β_q to be dispersed with equal energy in all spatial and temporal directions. We have empirically found that of two codes with equal mutual informations, the one satisfying the more stringent constraint gives lower error rates.

2. The solution to (10) subject to any of the constraints (i)–(iii) is highly nonunique: simply reordering the $\{A_q, B_q\}$ gives another solution, as does pre- or post-multiplying all the matrices by the same unitary matrix. However, there is also another source of nonuniqueness which is more subtle. Indeed, the transformation

$$\begin{array}{rcl} A'_q & = & \sum_{p=1}^{Q}\left(A_p \phi_{2p-1,2q-1} + B_p \phi_{2p,2q-1}\right) \\ B'_q & = & \sum_{p=1}^{Q}\left(A_p \phi_{2p-1,2q} + B_p \phi_{2p,2q}\right) \end{array} \qquad q = 1, \ldots Q \tag{11}$$

   where $\Phi = \{\phi_{i,j}\}_{i,j=1}^{2Q}$ is a real orthogonal matrix, preserves the mutual information. Judicious choice of Φ allows one to to improve the coding performance of a dispersion code $\{A_q, B_q\}$ without sacrificing its maximum mutual information.

3. The cost function $\frac{1}{2T}\mathrm{E} \log\det\left(I_{2NT} + \frac{\rho}{M}\mathcal{H}\mathcal{H}^t\right)$ can have local maxima. Nevertheless, we have been able to solve (10) with relative ease using gradient-based methods.

4. When $N \geq M$ and $Q = MT$ one solution to (10), subject to either constraints (i) or (ii), is given by the V-BLAST matrices. However, these are rank-one and therefore do not

satisfy constraint (iii). Using item 2 above, it is possible to obtain solutions subject to (iii). For $T = M$, one such set of matrices is given by

$$A'_{M(k-1)+l} = B'_{M(k-1)+l} = \frac{1}{\sqrt{M}} D^{k-1}\Pi^{l-1}, \quad k = 1, \dots, M, \; l = 1, \dots, M \tag{12}$$

where

$$D = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & e^{j\frac{2\pi}{M}} & 0 & \cdots \\ \vdots & & \ddots & \\ 0 & & & e^{j\frac{2\pi(M-1)}{M}} \end{bmatrix}, \qquad \Pi = \begin{bmatrix} 0 & \cdots & & 0 & 1 \\ 1 & 0 & \cdots & & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix}.$$

5. The design criterion (10) *depends explicitly on the number of receive antennas N*, both through the choice of Q and through the matrix $\mathcal{H}$. Hence, the optimal codes, for a given ρ, T, and M, are different for different N.

4 Examples of LD Codes and Performance

In this section we present simulations that compare the performance of LD codes to V-BLAST and orthogonal designs over a wide range of SNR's and various combinations of receive and transmit antennas.

LD vs. OD: $M = N = 2$, $R = 8$

We test the performance of the 2×2 orthogonal design when $M = N = 2$ at $R = 8$ versus the LD code given by (12) for $T = 2$ and $Q = 4$. The result can be seen in Figure 2 which clearly shows the superior performance of the LD code. To achieve $R = 8$, the orthogonal design needs to choose s_1 and s_2 from a 256-QAM constellation, while the LD code can choose from a 16-QAM constellation. We note that the orthogonal design achieves 7.47 bits/channel use mutual information at $\rho = 20$ dB, while the LD code achieves the full channel capacity of 11.28 bits/channel use. The decoding is ML, since this was feasible.

LD vs. OD: $M = 3$, $N = 1$, $R = 6$

We present a code for $M = 3$ transmit antennas and $N = 1$ receive antennas and compare it with the orthogonal design [8]

$$S = \sqrt{\frac{4}{3}} \begin{bmatrix} \alpha_1 + j\beta_1 & \alpha_2 + j\beta_2 & \alpha_3 + j\beta_3 \\ -\alpha_2 + j\beta_2 & \alpha_1 - j\beta_1 & 0 \\ -\alpha_3 + j\beta_3 & 0 & \alpha_1 - i\beta_1 \\ 0 & -\alpha_3 + j\beta_3 & \alpha_2 - i\beta_2 \end{bmatrix}. \tag{13}$$

It turns out that this achieves 5.13 bits/channel at $\rho = 20$ dB, whereas the channel capacity is 6.41 bits/channel use. To find a better LD code, we choose $Q = 4$ (recall that $N = 1$ and $T = 4$) and optimize (10) using gradient search to obtain

$$S = \begin{bmatrix} \alpha_1+\alpha_3+j[\frac{\beta_2+\beta_3}{\sqrt{2}}+\beta_4] & \frac{\alpha_2-\alpha_4}{\sqrt{2}}-j[\frac{\beta_1}{\sqrt{2}}+\frac{\beta_2-\beta_3}{2}] & 0 \\ \frac{-\alpha_2+\alpha_4}{\sqrt{2}}-j[\frac{\beta_1}{\sqrt{2}}+\frac{\beta_2-\beta_3}{2}] & \alpha_1-j\frac{\beta_2+\beta_3}{\sqrt{2}} & -\frac{\alpha_2+\alpha_4}{\sqrt{2}}+j[\frac{\beta_1}{\sqrt{2}}-\frac{\beta_2-\beta_3}{2}] \\ 0 & \frac{\alpha_2+\alpha_4}{\sqrt{2}}+j[\frac{\beta_1}{\sqrt{2}}-\frac{\beta_2-\beta_3}{2}] & \alpha_1-\alpha_3+j[\frac{\beta_2+\beta_3}{\sqrt{2}}-\beta_4] \\ \frac{\alpha_2-\alpha_4}{\sqrt{2}}+j[\frac{\beta_1}{\sqrt{2}}+\frac{\beta_2-\beta_3}{2}] & -\alpha_3+j\beta_4 & -\frac{\alpha_2+\alpha_4}{\sqrt{2}}+j[\frac{\beta_1}{\sqrt{2}}-\frac{\beta_2-\beta_3}{2}] \end{bmatrix}, \tag{14}$$

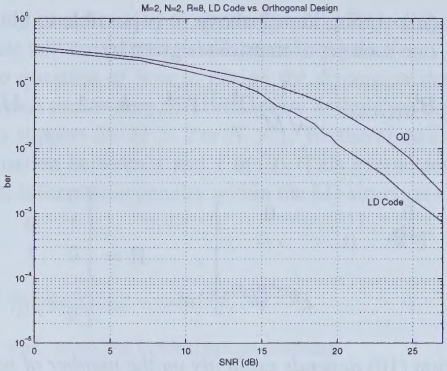


Figure 2: Bit error performance of the 2×2 orthogonal design versus the LD code given by (12) for $M = N = T = 2$ and $Q = 4$. The decoding is maximum likelihood.

which achieves 6.25 bits/channel use at $\rho = 20$ dB. Interestingly, S has orthogonal (but not orthonormal) columns.

Figure 3 shows the superior performance of the LD code. at rate 6. The entries of the orthogonal design are 256-QAM and of the LD code 64-QAM. The orthogonal design is ML decoded and the LD code is decoded using efficient nulling/cancelling [12].

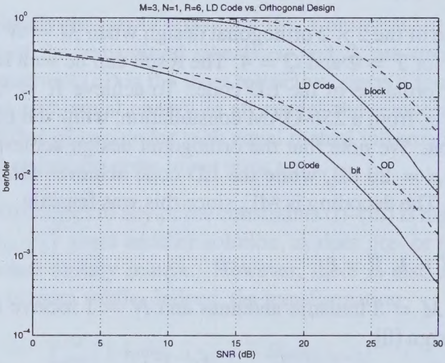


Figure 3: Block and bit error performance of orthogonal design (13) and the LD code (14) for $M = 3$ antennas at rate is $R = 6$. The decoding for (13) is ML and for (14) efficient nulling/cancelling [12].

LD vs. V-BLAST: $M = 4, 8, N = 4, R = 16$

Figure 4 compares an LD code for $M = 8$, $N = 4$ with $M = 4$, $N = 4$ V-BLAST at rate $R = 16$. The $M = N = 4$ and $M = 8$, $N = 4$ channel capacities at 20 dB are 22.14 and 24.94 bits/channel use, respectively, whereas the optimized LD code achieves 23.10 bits/channel use with $T = 8$ and $Q = 32$.

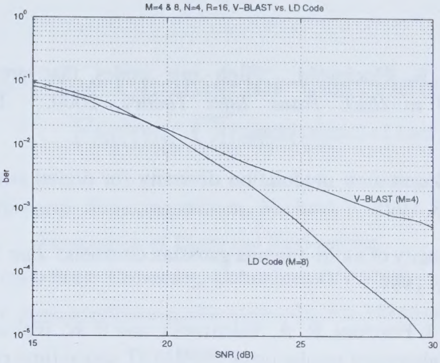


Figure 4: Bit error performance of V-BLAST for $M = 4$ and $N = 4$ (??) and an LD code for $M = 8$ and $N = 4$ with $Q = 32$ at rate is $R = 16$ with nulling/cancelling decoding.

4.1 Table of Mutual Informations

Table 1 summarizes the mutual informations of some LD codes that we generated and the actual channel capacities at $\rho = 20$ dB. As can be observed, $C_{\mathrm{LD}}(\rho, T, M, N)$ is very close to $C(\rho, M, N)$; there is little penalty in the linear structure of the dispersion codes.

(T,M,N)	$C_{LD}(\rho, T, M, N)$	$C(\rho, M, N)$	Reference
(2,2,1)	6.28	6.28	Eq. (4)
(4,3,1)	6.25	6.41	Eq. (14), Fig. 3
(6,3,1)	6.28	6.41	
(4,3,2)	11.63	12.14	
(4,4,1)	6.34	6.47	
(4,4,2)	11.99	12.49	
(8,8,4)	23.10	24.94	Fig. 4
$(T = M, M, N \geq M)$	$C(\rho, M, N)$	*	Eq. (12), Figs. 2

Table 1: Mutual information $C_{LD}(\rho, T, M, N)$ obtained via gradient-ascent optimization of the cost function (10), compared to the actual channel capacity $C(\rho, M, N)$.

5 Conclusion

The linear dispersion codes we have introduced are simple to encode and decode, apply to any combination of transmit and receive antennas, subsume as special cases many earlier space-time transmission schemes, and satisfy an information-theoretic optimality property. Our simulations confirm that information-theoretic optimality translates to good code performance in raw block and bit error—but it is also important for attaining channel capacity. The codes are designed to be linear while having little (if any) penalty in mutual information, and additional channel coding across $s_1, \ldots, s_Q$ can be combined with an LD code to attain most (if not all) of the channel capacity.

References

[1] B. Hassibi and B. Hochwald, "High rate codes that are linear in space and time," *submitted to IEEE Trans. Info. Theory*, 2000. Download available at `http://mars.bell-labs.com`.

[2] A. Wittneben, "Basestation modulation diversity for digital simulcast," in *Proc. IEEE VTC*, pp. 848–853, 1991.

[3] I. E. Telatar, "Capacity of multi-antenna gaussian channels," *Eur. Trans. Telecom.*, vol. 10, pp. 585–595, Nov. 1999.

[4] G. D. Golden, G. J. Foschini, R. A. Valenzuela, and P. W. Wolniansky, "Detection algorithm and initial laboratory results using V-BLAST space-time communication architecture," *Electronic Letters*, vol. 17, Nov. 1998.

[5] J.-C. Guey, M. P. Fitz, M. R. Bell, and W.-Y. Kuo, "Signal design for transmitter diversity wireless communication systems over rayleigh fading channels," in *Proc. IEEE VTC*, pp. 136–140, 1996.

[6] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication: Performance criterion and code construction," *IEEE Trans. Info. Theory*, vol. 44, pp. 744–765, 1998.

[7] S. M. Alamouti, "A simple transmitter diversity scheme for wireless communications," *IEEE J. Sel. Area Comm.*, pp. 1451–1458, Oct. 1998.

[8] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Trans. Info. Theory.*, vol. 45, pp. 1456–1467, July 1999.

[9] B. Hochwald and W. Sweldens, "Differential unitary space time modulation," tech. rep., Bell Laboratories, Lucent Technologies, Mar. 1999. To appear in *IEEE Trans. Comm.*. Download available at `http://mars.bell-labs.com`.

[10] B. Hughes, "Differential space-time modulation," *submitted to IEEE Trans. Info. Theory*, 1999.

[11] B. Hassibi, A. Shokrollahi, B. Hochwald, and W. Sweldens, "Representation theory for high-rate multiple-antenna code design," *submitted to IEEE Trans. Info. Theory*, 2000. Download available at `http://mars.bell-labs.com`.

[12] B. Hassibi, "An efficient square-root algorithm for blast," *submitted to IEEE Trans. Sig. Proc.*, 1999. Download available at `http://mars.bell-labs.com`.

[13] O. Damen, A. Chkeif, and J.-C. Belfiore, "Lattice code decoder for space-time codes," *IEEE Comm. Let.*, pp. 161–163, May 2000.

[14] G. J. Foschini, "Layered space-time architecture for wireless communication in a fading environment when using multi-element antennas," *Bell Labs. Tech. J.*, vol. 1, no. 2, pp. 41–59, 1996.

Robust Space-Time Coding for Correlated Rayleigh Fading Channels

Siwaruk Siwamogsatham, Michael P. Fitz
Department of Electrical Engineering
Ohio State University, Columbus, OH 43210
Email: siwamogs@ee.eng.ohio-state.edu, fitz.7@osu.edu

Abstract

The traditional design criteria for constructing space-time codes are not suitable for Rayleigh fading channels with large space-time correlations. In this paper, we propose a coding scheme and design criteria for constructing space-time codes, namely smart-robust codes, that achieve robust performance in practical channels. We present a simple class of the smart-robust codes that combines the multiple trellis-coded modulation (M-TCM) scheme with a space-time constellation. Simulation results demonstrate the proposed codes deliver robust performance over a wide range of channel conditions.

1 Introduction

Space-time coding is a promising method to achieve an improved performance for wireless communications in Rayleigh fading environments. By exploiting the benefits of transmit and receive antenna diversity, space-time coding can enable higher throughput or more reliable transmissions. The standard design criteria for constructing good space-time codes were proposed in [1, 2] for quasi-static and fast fading channels. These criteria have been commonly used in the literature to construct various classes of space-time codes based on trellis-coded modulation (TCM) [1, 3, 4, 5], block coding [2, 6, 7, 8, 9], concatenated coding [10, 11, 12, 13], and etc..

So far, the existing space-time coding schemes were designed for the ideal space-time channels where the space-time correlations among the signal paths were negligible. In practice, however, it is often difficult to satisfy the ideal condition, and the existing space-time codes may suffer greatly. Recent studies [14] pointed out that the performance of the smart-greedy codes proposed [1] degrade drastically in Rayleigh fading channels with certain space-time correlations. In some cases where channels have large correlations across space and time, the smart-greedy codes in the seminal paper [1] perform surprisingly poor although the channels have small temporal correlations for each spatial channel (rich time diversity) and small spatial correlations for each time instant (rich space diversity). The undesirable performance degradation is caused by the fact that the standard design criteria are no longer optimal for the non-ideal space-time channels. Up to the current time, no one has addressed a design of space-time codes that achieve robust performance in correlated Rayleigh fading channels.

In this paper, we propose a robust space-time coding scheme for Rayleigh fading channels with space-time correlations. Our design is simple, yet robust and powerful. The key component of our scheme is the use group-wise interleaving instead of the conventional symbol-by-symbol interleaving, which allows us to assume that L_t consecutive received samples experience constant channels. With this assumption, it is possible to formulate the design criteria that are not channel-dependent. We present a few examples of simple multiple trellis-coded modulation (M-TCM) construction based on the proposed criteria. Our smart-robust space-time codes are robust against space-time correlations of the channels, and more powerful than the original smart-greedy codes. It is quite intriguing that the robustness is achieved without compromising performance. More code designs are left as future work.

2 Scheme Formulation

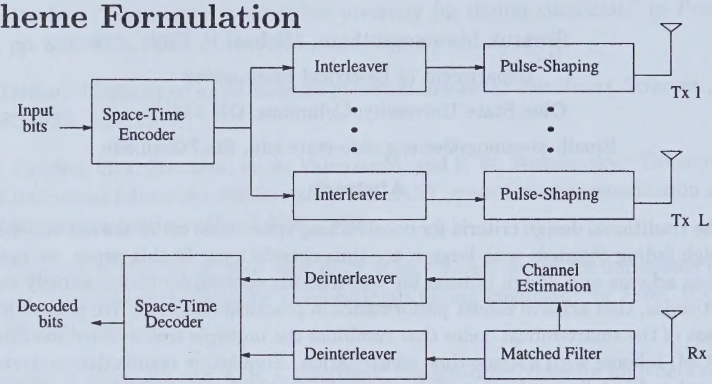


Figure 1: A block diagram of a typical transmit diversity system

Basically, the only difference of the proposed smart-robust scheme from a typical space-time coding scheme is the use of group-wise interleaver instead of a conventional symbol-by-symbol interleaver. The group-wise interleaver operates in a unit of L_t symbols so that channels can be assumed constant over each group of L_t received samples. With this assumption that $C_i(k+l) = C_i(k)$ for $0 \leq l \leq L_t - 1$, the $N_c \times 1$ vector of the received samples has a nice form:

$$\vec{Q} = \sqrt{E_s}\mathbf{D}\vec{C} + \vec{N} \tag{1}$$

where the codeword matrix $\mathbf{D}$ here is an $N_c \times N_c$ matrix formed by $L_t \times L_t$ submatrices along the diagonal entries, i.e., for $N_c' = N_c/L_t$,

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}(1) & 0 & \cdots & 0 \\ 0 & \mathbf{D}(2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \mathbf{D}(N_c') \end{bmatrix} \tag{2}$$

where

$$\mathbf{D}(k) = \begin{bmatrix} D_1\left((k\text{-}1)L_t\text{+}1\right) & D_2\left((k\text{-}1)L_t\text{+}1\right) & \cdots & D_{L_t}\left((k\text{-}1)L_t\text{+}1\right) \\ D_1\left((k\text{-}1)L_t\text{+}2\right) & D_2\left((k\text{-}1)L_t\text{+}2\right) & \cdots & D_{L_t}\left((k\text{-}1)L_t\text{+}2\right) \\ \vdots & \vdots & \ddots & \vdots \\ D_1(kL_t) & D_2(kL_t) & \cdots & D_{L_t}(kL_t) \end{bmatrix}$$

and the channel distortion vector

$$\vec{C} = \left[C_1\left(\Pi^{-1}(1)\right), C_2\left(\Pi^{-1}(1)\right), \ldots, C_{L_t}\left(\Pi^{-1}(1)\right), C_1\left(\Pi^{-1}(L_t\text{+}1)\right), \ldots, C_{L_t}\left(\Pi^{-1}(N_c\text{-}L_t\text{+}1)\right)\right]^T$$

is an $N_c \times 1$ vector for this case, as opposed to the $N_c L_t \times 1$ channel distortion vector in the conventional case. Note that the codeword matrix here is a square matrix and has the same size as the covariance matrix of $\vec{C}$. This property enables us to formulate code design criteria that are not channel-dependent.

3 Smart-Robust Design Criteria

Typically, performance of a coding scheme is analyzed using the pairwise error probability (PWEP), the probability that a decoder erroneously chooses a codeword $\mathbf{D} = \mathbf{d}_\beta$ over a transmitted codeword $\mathbf{D} = \mathbf{d}_\alpha$. For a coherent detection scheme with perfect channel state information (CSI) available at the maximum likelihood decoder, according to [14], this probability is well bounded by

$$P(\mathbf{d}_\alpha \to \mathbf{d}_\beta) \leq \begin{pmatrix} 2\Delta_H - 1 \\ \Delta_H - 1 \end{pmatrix} \prod_{i=1}^{\Delta_H} \left(\frac{N_0}{\lambda_i E_s} \right) \tag{3}$$

where λ_i, for $1 \leq i \leq \Delta_H$, are the non-zero eigenvalues of the signal matrix

$$\mathbf{C}_s = \mathbf{B}(\alpha,\beta)\mathbf{C}_c\mathbf{B}(\alpha,\beta)^H \tag{4}$$

where $\mathbf{B}(\alpha,\beta) = \mathbf{d}_\alpha - \mathbf{d}_\beta$ denotes the codeword difference matrix, and $\mathbf{C}_c$ denotes the covariance matrix of the channel distortion $\vec{C}$. For the proposed scheme, the codeword different matrix $\mathbf{B}(\alpha,\beta)$ and the channel covariance matrix $\mathbf{C}_c$ are both $N_c \times N_c$ matrices. It is observed that the codeword matrix $\mathbf{B}(\alpha,\beta)$ generally has many zero rows and columns which may be excluded from the analysis. So let's define $\mathbf{B}'(\alpha,\beta)$ to be the reduced codeword difference matrix in which only non-zero rows and columns are included, and define $\mathbf{C}'_c$ to be the covariance matrix of the relevant channel distortion coefficients such that if the i^{th} row and column of $\mathbf{B}(\alpha,\beta)$ are removed to form $\mathbf{B}'(\alpha,\beta)$ then the corresponding i^{th} row and column of $\mathbf{C}_c$ are removed to form $\mathbf{C}'_c$. Note that the non-zero eigenvalues of the signal matrix in (4) are the same as the nonzero eigenvalues of the following signal matrix

$$\mathbf{C}'_s = \mathbf{B}'(\alpha,\beta)^H\mathbf{B}'(\alpha,\beta)\mathbf{C}'_c. \tag{5}$$

In most cases with a non-zero Doppler frequency, the covariance matrix $\mathbf{C}'_c$ generally has full rank, and if we make $\mathbf{B}'(\alpha,\beta)^H\mathbf{B}'(\alpha,\beta)$ to have full rank, then the signal matrix $\mathbf{C}'_s$ has full rank and the PWEP can be bounded by

$$P(\mathbf{d}_\alpha \to \mathbf{d}_\beta) \leq \begin{pmatrix} 2\Delta_H - 1 \\ \Delta_H - 1 \end{pmatrix} \left(\frac{N_0}{E_s} \right)^{\Delta_H} \frac{1}{\det(\mathbf{C}'_s)} \tag{6}$$

in which the determinant of $\mathbf{C}'_s$ can be decomposed as

$$\det(\mathbf{C}'_s) = \det\left(\mathbf{B}'(\alpha,\beta)^H\mathbf{B}'(\alpha,\beta)\right)\det\left(\mathbf{C}'_c\right). \tag{7}$$

Note (7) is only true when $\mathbf{B}'(\alpha,\beta)^H\mathbf{B}'(\alpha,\beta)$ and $\mathbf{C}'_s$ are of full rank and the same size. The group-wise interleaving structure here-in proposed and proper code designs guarantee these conditions hold. It can be seen that the performance of a smart-robust code is inversely proportional to $\det\left(\mathbf{B}'(\alpha,\beta)^H\mathbf{B}'(\alpha,\beta)\right)$ and the determinant of $\mathbf{C}'_c$ may be viewed as the loss due to channel correlations. Note that with the proposed formulation, the choices of good codes do not depend on the channels.

Based on the above analysis, we arrive at the following universal design criteria for smart and robust space-time codes:

- ***The Rank Criterion:*** The following matrix

  $$\Lambda(\alpha,\beta) = \mathbf{B}'(\alpha,\beta)^H\mathbf{B}'(\alpha,\beta) \tag{8}$$

  shall have full rank, and the minimum rank of $\Lambda(\alpha,\beta)$ for any two distinct codeword $\mathbf{d}_\alpha$ and $\mathbf{d}_\beta$ determines maximum diversity advantage of a space-time code for the proposed scheme. Denote $\text{rank}(\Lambda(\alpha,\beta)) = \Delta_H(\alpha,\beta)$.

- ***The Determinant Criterion:*** The determinant of $\Lambda(\alpha,\beta)$,

$$\Delta_p(\alpha,\beta) = \det\{\Lambda(\alpha,\beta)\} \tag{9}$$

must be maximized over every pair of distinct codeword $\mathbf{d}_\alpha$ and $\mathbf{d}_\beta$ in order to achieve the most coding advantage.

For simplicity in notation for the sequel $\Delta_{\text{H, ideal}}$ and $\Delta_{\text{p, ideal}}$ denote the minimum diversity advantage and coding gain for ideal fading respectively. The notation for static fading will be $\Delta_{\text{H, static}}$ and $\Delta_{\text{p, static}}$.

It is worthwhile to remark that the robust design criteria derived above are also valid when the group-wise interleaver operates in a unit of $L > L_t$. In such a case, the reduced codeword difference matrix $\mathbf{B}'(\alpha,\beta)$ has more rows than columns, and hence $\Lambda(\alpha,\beta) = \mathbf{B}'(\alpha,\beta)^H\mathbf{B}'(\alpha,\beta)$ can still be made to satisfy the full rank requirement. However, a group-wise interleaver with a larger number of symbols than antennas ($L > L_t$) may offer less temporal diversity.

4 Code Examples

The criteria we proposed can be used to design different classes of space-time codes. In this section, we present a simple class of the smart-robust codes that combines the multiple trellis-coded modulation (M-TCM) scheme with a space-time constellation or a space-time block code, e.g. [2, 6, 7, 8, 9]. Here, we use the orthogonal space-time constellations [6, 7]. The design technique is explained by examples below.

Example 1: In this example, we will construct a smart-robust space-time code with a 0.5-bit/symbol transmission rate for a system with 2 transmit antennas using BPSK constellation. The goal is to construct a smart-robust code with equal or better performance measures than the original BPSK smart-greedy space-time code [1] with the same complexity. The smart-greedy code achieves $\Delta_{\text{H, ideal}} = 4$ for the ideal fading cases, and $\Delta_{\text{H, static}} = 2$ for the quasi-static fading case.

In this example, the Alamouti's space-time constellation [6] is combined with an M-TCM with multiplicity 2 as shown in Fig. 10. Each transition of the M-TCM produces two symbols which will eventually be transmitted via the Alamouti's signal modulation. For example, if the branch labeled s_0s_1 is chosen at time k, then s_0 and $-s_1^*$ are transmitted at time k and $k+1$ via antenna 1, and s_1 and s_0^* are transmitted at time k and $k+1$ via antenna 2, respectively. Note that $(\bullet)^*$ denotes the complex conjugate operation. Let's define the equivalent space-time Euclidean distance between a given pair of branch labels at time k, $\vec{s}(k) = [s_0, s_1]$ and $\vec{w}(k) = [w_0, w_1]$ as

$$\begin{aligned}\tilde{d}(\vec{s}(k),\vec{w}(k)) &= \det\left(\begin{bmatrix} s_0 - w_0 & s_1 - w_1 \\ w_1^* - s_1^* & s_0^* - w_0^* \end{bmatrix}^H \begin{bmatrix} s_0 - w_0 & s_1 - w_1 \\ w_1^* - s_1^* & s_0^* - w_0^* \end{bmatrix}\right) \\ &= \left(|s_0 - w_0|^2 + |s_1 - w_1|^2\right)^2.\end{aligned} \tag{10}$$

Then, the determinant criterion in (9) is simply the product of the equivalent space-time Euclidean distance taken over the set of time instances $1 \le k \le N_c$ in which the branch labels, $\vec{s}(k)$ and $\vec{w}(k)$, are distinct, i.e.,

$$\Delta_p = \prod_{k \in v} \tilde{d}(\vec{s}(k),\vec{w}(k)) \tag{11}$$

where v denotes the the set of time instances in which the branch labels are distinct, for a given pair of code sequences. By noting analog between the equivalent space-time Euclidean distance and the conventional squared Euclidean distance, and that the standard criterion for designing an M-TCM code in AWGN or fading channels and (11) are quite similar, we adopt the standard set partitioning principal in [15] for designing our code. One must be cautious though that our design goal is to maximize the product of the equivalent space-time Euclidean distance and the length of the minimum distance path, while that in [15] was to maximize the sum of the squared Euclidean distance. Nonetheless, the classic set partitioning works well for our design.

Here, we form the signal sets $A_0 = [00, 11]$ and $A_1 = [01, 10]$ such that the elements from the same set have the minimum equivalent space-time Euclidean distance equal to $(8)^2$ and the elements from different sets have the minimum equivalent space-time Euclidean distance equal to $(4)^2$, i.e., $\tilde{d}_{\min}(A_0) = \tilde{d}_{\min}(A_1) = (8)^2$ and $\tilde{d}_{\min}(A_0, A_1) = (4)^2$. We construct the code in Fig. 10 by assigning the elements from A_0 to the first state and the elements from A_1 to the second state. In this example, the BPSK smart-robust code has $\Delta_{\text{H, ideal}} = 4$ with $\Delta_{\text{p, ideal}} = (32)^2$ for non-quasi-static cases, and $\Delta_{\text{H, static}} = 2$ with $\Delta_{\text{p, static}} = 144$ for the case of quasi-static fading. Compared with the original BPSK smart-greedy space-time code which has $\Delta_{\text{H, ideal}} = 4$ with $\Delta_{\text{p, ideal}} = (32)^2$ for the ideal fading case, and $\Delta_{\text{H, ideal}} = 2$ with $\Delta_{\text{p, ideal}} = 80$ for the quasi-static case, the new code should perform slightly better in the quasi-static fading case, and equally in the ideal fading case. The numerical results are given in the simulations section where it is shown that our code performs better than the smart-greedy code in a wide range of channel conditions, especially when the channels have large space-time

Example 2: In this example, we will construct a space-time code with a 1-bit/symbol transmission rate for a system with 2 transmit antennas with QPSK constellations. The constructed code has slightly greater decoding complexity than the QPSK smart-greedy code, but achieves substantially more powerful performance.

In this example, we employ the combined M-TCM and Alamouti's signals technique as suggested earlier. We adopt the following set partitions: $A_0 = [00, 02, 22, 20]$, $A_1 = [12, 10, 30, 32]$, $A_2 = [33, 31, 11, 13]$, and $A_3 = [21, 23, 03, 01]$. These sets have the following minimum equivalent Euclidean distances: $\tilde{d}_{\min}(A_i, A_j) = (4)^2$, for $i, j = 0, 1, 2, 3$. We assign the elements of A_0, A_1, A_2, and A_3 to the paths leaving state 0, 1, 2, and 3, respectively. Fig. 12 shows the proposed 4-state 1-bit/symbol QPSK smart-robust code. We have optimized this construction in terms of the worst-case Δ_p. The following design rules are employed. First, the distance between adjacent paths leaving a given node is equal to $(4)^2$, and distance between non-adjacent paths leaving a given node is equal to $(8)^2$. Second, the distance between adjacent paths merging to a given node is equal to $(6)^2$, and distance between non-adjacent paths merging to a given node is equal to $(4)^2$. Third, the minimum distance between any two paths is equal to $(2)^2$. For the quasi-static fading case, this code achieves $\Delta_{\text{H, static}} = 2$ with $\Delta_{\text{p, static}} = (10)^2$, and $\Delta_{\text{H, ideal}} = 4$ with $\Delta_{\text{p, ideal}} = (24)^2$. Compared with the original QPSK smart-greedy code which achieves $\Delta_{\text{H, static}} = 2$ with $\Delta_{\text{p, static}} = 80$ and $\Delta_{\text{H, ideal}} = 3$ with $\Delta_{\text{p, ideal}} = 16$, it can be predicted that our code should considerably outperform the QPSK smart-greedy code.

Example 3: In this example, we demonstrate a technique to design a more powerful smart-robust code for 2-transmit antenna systems by employing an M-TCM encoder with multiplicity greater than 2, but still using a 2-symbol group-wise interleaver. Fig. 11 depicts a 4-state, 0.5-bit/symbol, BPSK smart-robust code. Here, we use an M-TCM encoder with multiplicity equal to 4. In this case, the distance between a pair of four symbols at time k, $\vec{s}(k) = [s_0, s_1, s_2, s_3]$ and $\vec{w}(k) = [w_0, w_1, w_0, w_1]$ is given by $\tilde{d}(\vec{s}(k), \vec{w}(k)) = \left(|s_0 - w_0|^2 + |s_1 - w_1|^2\right)^2 \left(|s_2 - w_2|^2 + |s_3 - w_3|^2\right)^2$. For the quasi-static fading case, this

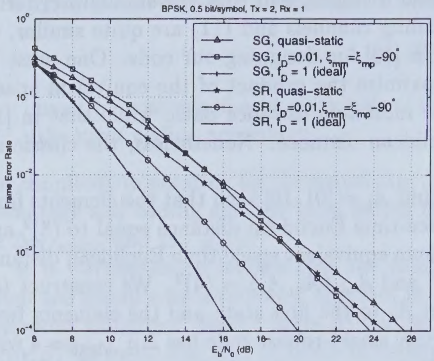


Figure 2: Frame error rate vs. SNR for the 2-state BPSK codes

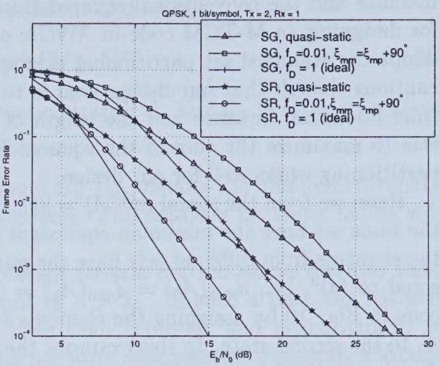


Figure 3: Frame error rate vs. SNR for the QPSK codes

code achieves $\Delta_{\text{H, static}} = 2$ with $\Delta_{\text{p, static}} = (20)^2$, corresponding to $10\log_{10}(20/12) = 2.22$ dB in the minimum coding gain over the 2-state BPSK smart-robust code (though, the average performance gain may be smaller.) For non-zero Doppler frequency cases, this code achieves $\Delta_{\text{H, ideal}} = 8$ with $\Delta_{\text{p, ideal}} = (512)^2$.

Example 4: In this example, we will construct a smart-robust space-time code with 0.5-bit/symbol transmission rate for a system with 4 (or 3) transmit antennas using BPSK constellation. Here, we combine the M-TCM scheme (with multiplicity of 4) with the 4×4 orthogonal BPSK block code[7]. In this case, the equivalent space-time Euclidean distance between a given pair of four symbols at time k, $\vec{s}(k) = [s_0, s_1, s_2, s_3]$ and $\vec{w}(k) = [w_0, w_1, w_0, w_1]$ is given as $\tilde{d}(\vec{s}(k), \vec{w}(k)) = (|s_0 - w_0|^2 + |s_1 - w_1|^2 + |s_2 - w_2|^2 + |s_3 - w_3|^2)^4$. For the quasi-static fading case, this code achieves $\Delta_{\text{H, static}} = 4$ with $\Delta_{\text{p, static}} = (20)^4$, and $\Delta_{\text{H, ideal}} = 8$ with $\Delta_{\text{p, ideal}} = (96)^4$. In addition, the smart-robust code in Fig. 14 can also be applicable for a system with 3 transmit antennas, by utilizing just 3 out of 4 columns of the block. In this case, the code achieves $\Delta_{\text{H, static}} = 3$ with $\Delta_{\text{p, static}} = (20)^3$, and $\Delta_{\text{H, ideal}} = 6$ with $\Delta_{\text{p, ideal}} = (96)^3$.

5 Simulation Results

In this section, we investigate performance of the smart-robust codes and the smart-greedy codes in variety of channel conditions. The main objective of the following simulation results is to demonstrate the robustness of the proposed designs. In addition, these simulation results will show that the performance of the smart-greedy scheme is greatly dependent on channel correlations. The performance can be quite poor although the channel conditions are not necessarily bad. The interaction of the smart-greedy code and channel produce poor distance characteristics. In contrast, the smart-robust codes achieve superior and dependable performance in a wide range of channel conditions. One can expect good performance from the smart-robust codes when the channels are decent, and reasonable degradation in performance when the channels are correlated.

We analyze performance of the codes via the frame error rate simulations. For the robustness studies, we simulate the frame error rate of the BPSK (2-state) and QPSK smart-robust codes, as well as the BPSK and QPSK smart-greedy codes for 2-transmit antenna systems as a function of different channel parameters. In the simulations, a frame is formed

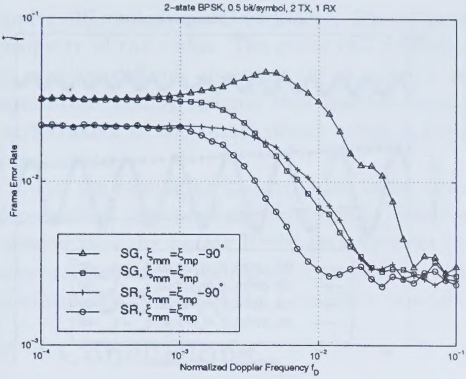


Figure 4: Frame error rate vs. Doppler frequency for the 2-state BPSK codes: $d_{sp} = 30\lambda$, $a_\lambda = 25\lambda$, $d_\lambda = 1500\lambda$, $\xi_{mp} = 30°$, $\frac{E_s}{N_0} = 10$ dB.

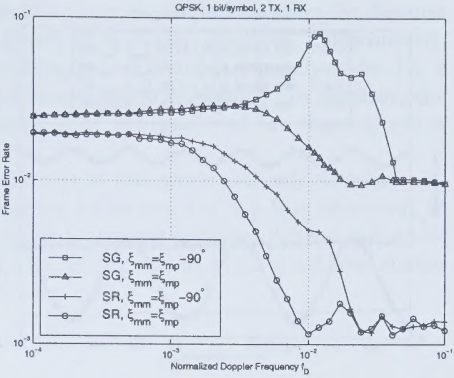


Figure 5: Frame error rate vs. Doppler frequency for the QPSK codes: $d_{sp} = 30\lambda$, $a_\lambda = 25\lambda$, $d_\lambda = 1500\lambda$, and $\xi_{mp} = 30°$.

by 126 symbols plus additional symbols to close the trellis (two symbols for most cases except the QPSK smart-greedy code that requires 3 symbols to close the trellis.) A 16×8 symbol-by-symbol block interleaver is used for the smart-greedy scheme, and a $[2 \times 16] \times 4$ group-wise interleaver (which effectively separates each 2-symbol group by at least 2×16 symbols) is used for our scheme. Note that for a fixed delay constraint, a group-wise interleaver that operates in a unit of L_t symbols allows L_t times larger group spacing than symbol spacing allowable by a symbol-by-symbol interleaver. The channels are simulated by the space-time channel model given in [16] (see Fig. 15). Note that this model produces large correlations when $\xi_{mm} = \xi_{mp} \pm 90°$.

Fig. 8 depicts the performance of the smart-robust codes and the smart-greedy codes as a function of mobile moving direction (ξ_{mm}), with $d_\lambda = 1500\lambda$, $d_{sp} = 29.2\lambda$, $a_\lambda = 25\lambda$, $f_D = 0.0175$, and $\xi_{mp} = 30°$. Since the channel model produces large space-time correlations when $\xi_{mm} = \xi_{mp} \pm 90°$, we can notice that degradation in performance occurs at these points. We observe that the performance of the BPSK smart-greedy code degrades severely at $\xi_{mm} = \xi_{mp} - 90°$, but degrades little at $\xi_{mm} = \xi_{mp} + 90°$, although the magnitude of the space-time correlations are large in both cases. The large space-time correlations may or may not hurt performance of the smart-greedy codes, depending on how the codes interact with the channel correlations, and the degradation can be quite extreme when the channel correlations are destructive to the distance property of the smart-greedy codes. We also note the similar trend in performance of the QPSK smart-greedy code, except the more severe degradation happens at $\xi_{mm} = \xi_{mp} + 90°$. In contrast, the smart-robust codes have no problem achieving decent performance in such cases. Note also that, the proposed QPSK smart-robust code is much more powerful than the QPSK smart-greedy code. It performs better, even with 5 dB lower in SNR.

Fig. 2 to Fig. 5 show that when the space-time correlations are small, the smart-greedy codes perform as expected, i.e., performance improves as the Doppler frequency increases. However, when the space-time correlations are large, this desirable characteristic may not hold. For example, with the given parameters ($d_{sp} = 30\lambda$, $a_\lambda = 25\lambda$, $d_\lambda = 1000\lambda$, $\xi_{mp} = 30°$, $\xi_{mm}=\xi_{mp} + 90°$, and f_D=0.01 with a 16×8 symbol-by-symbol block interleaver), the

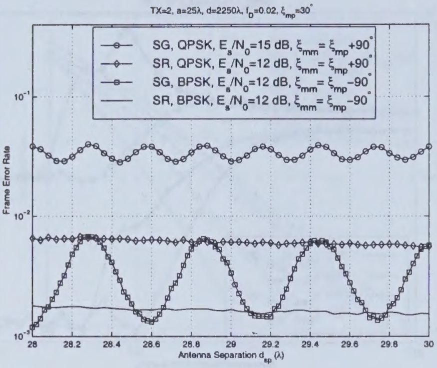


Figure 6: Frame error rate vs. antenna separations (d_{sp}) : $d_\lambda = 2250\lambda$, $a_\lambda = 25\lambda$, $\xi_{mp} = 30°$, $f_D = 0.02$.

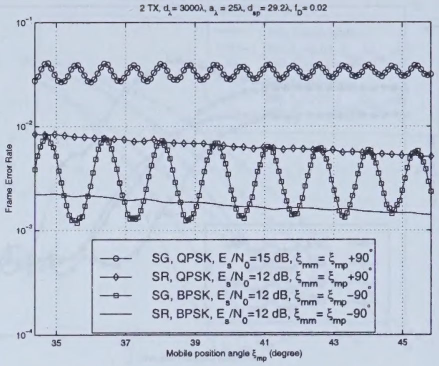


Figure 7: Frame error rate vs. mobile position angle (ξ_{mp}) : $d_{sp} = 29.2\lambda$, $d_\lambda = 3000\lambda$, $a_\lambda = 25\lambda$, $f_D = 0.02$.

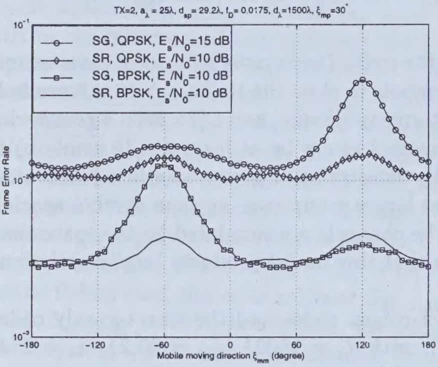


Figure 8: Frame error rate vs. mobile moving direction ξ_{mm}: $d_\lambda = 1500\lambda$, $d_{sp} = 29.2\lambda$, $a_\lambda = 25\lambda$, $f_D = 0.0175$, and $\xi_{mp} = 30°$.

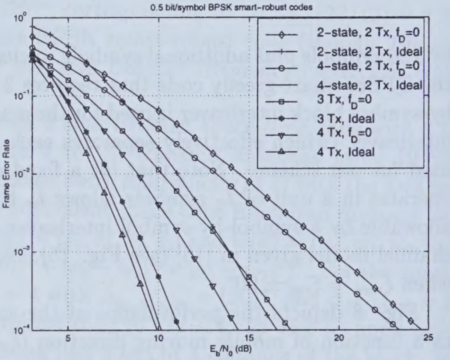


Figure 9: Frame error rate vs. SNR for 0.5-bit/symbol smart-robust BPSK codes

channel model produces large correlations across space and time (i.e., $|R_{c_1c_2}(i, d_{sp})| < 0.98$), yet small correlations for each channel in time (i.e. $|R_{c_1c_2}(16 \times i, 0)| < 0.78, i \neq 0$) and little correlations in space at the same time (i.e. $|R_{c_1c_2}(0, d_{sp})| < 0.05$); nonetheless, the QPSK smart-greedy code performs even worst than its performance in the quasi-static fading channels. In contrast, the smart-robust codes have no problem achieving decent performance in such channel conditions. Moreover, the performance of the smart-greedy codes can be quite sensitive to placement of antennas. and position angles of the mobile. Fig. 6 shows that the performance of the smart-greedy codes can alter drastically with a fractional drift in antenna separations, even when the antennas are spaced quite far apart ($d_{sp} > 28\lambda$). Furthermore, Fig. 7 shows that the performance of the smart-greedy codes is also sensitive to position angles of the mobile (ξ_{mp}), in the similar manner.

It should be evident by now that the performance of the smart-greedy codes greatly depends on channel correlations. The smart-greedy codes may or may not perform well under different channel conditions, depending on how channel correlations effect the distance property of the codes. The codes can perform poorly even in good channels. Consequently, it may be difficult to guarantee or predict performance of the smart-greedy codes (or a space-time coding scheme that employs symbol-by-symbol interleaving.) In contrast, the performance of the smart-robust codes is determined by the quality of the channel, and is predictable.

For the remaining smart-robust codes constructed in this paper, we only simulate performance of the codes as a function of signal-to-noise ratios (see Fig. 9.) It is interesting to observe that the 4-state BPSK smart-robust code for 2 transmit antennas delivers noticeable improvement in performance, especially for the ideal fading case in which it achieves similar performance to that of the 4-transmit antenna code.

6 Conclusions

The space-time codes that were derived from the traditional scheme and design criteria do not guarantee robust performance in Rayleigh fading channels with space-time correlations. In this paper, we have proposed a coding scheme and the design criteria for constructing space-time codes that achieve superior and robust performance. The robust scheme is formulated simply by using a group-wise interleaver instead of the symbol-by-symbol interleaver in the traditional scheme. Many classes of space-time codes can be constructed based on the proposed scheme and design criteria. A simple class of smart-robust codes that combines the multiple trellis-coded modulation (M-TCM) scheme with a space-time constellation was presented in this paper. The proposed codes delivered robust performance over practical channels. More code designs are left as future work.

References

[1] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high rate wireless communication: Performance and code construction," *IEEE Trans. Inform. Theory*, pp. 744–765, March 1998.

[2] J.-C. Guey, M. P. Fitz, M. R. Bell, and W.-Y. Kuo, "Signal design for transmitter diversity wireless systems over Rayleigh fading channels," *IEEE Trans. Commun.*, pp. 527–537, April 1999.

[3] J. Grimm, M. P. Fitz, and J. V. Krogmeier, "Further results on space-time codes for Rayleigh fading," *Proc. Allerton Conferrence*, pp. 391–400, September 1998.

[4] R. S. Blum, "Analytical tools for the design of space-time convolutional codes," submitted to *IEEE Trans. Inform. Theory*, Febuary 2000.

[5] S. Baro and G. Bauch. A. Hansmann, "Improved codes for space-time trellis-coded modulation," *IEEE Commun. Letters*, pp. 20–22, January 2000.

[6] S. M. Alamouti, "A simple transmit diversity technique for wireless communications," *IEEE J. Select. Commun.*, pp. 1451–1458, October 1998.

[7] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Trans. Information Theory*, vol. 45, pp. 1456–1467, July 1999.

[8] O. Tirkkonen and A. Hottinen, "Square-matrix embeddable space-time block codes for complex signal constellations," submitted to IEEE Trans. Inform. Theory.

[9] A. R. Hammons and H. E. Gamal, "On the theory of space-time codes for PSK modulation," *IEEE Trans. Information Theory*, pp. 524–542, March 2000.

[10] A. Yongacoglu and M. Siala, "Space-time codes for fading channels," *IEEE Vehicular Technology Conference*, pp. 2495–2499, 1999.

[11] G. Bauch, "Concatenation of space-time block codes and turbo-TCM," *IEEE Int. Conference on Commun.*, pp. 1202–1206, June 1999.

[12] A. S. Stefanov and T. M. Duman, "Turbo coded modulation for wireless communications with antenna diversity," *IEEE Vehicular Technology Conference*, pp. 1565–1569, September 1999.

[13] Y. Liu, M. P. Fitz, and O. Y. Takeshita, "QPSK space-time turbo codes," *IEEE International Conference on Communications*, pp. 292–296, June 2000.

[14] M. P. Fitz, J. Grimm, and S. Siwamogsatham, "A new view of performance analysis techniques in correlated Rayleigh fading," *WCNC*, pp. 139–144, September 1999.

[15] D. Divsalar and M. K. Simon, "Multiple trellis-coded modulation (M-TCM)," *IEEE Trans. Commun.*, vol. 36, no. 4, pp. 410–419, April 1988.

[16] T.-A. Chen, W.-Y. Kuo, M. P. Fitz, and M. D. Zoltowski, "A space-time model for frequency nonselective Rayleigh fading channels," *IEEE J. Select. Commun.*, pp. 1175–1190, July 2000.

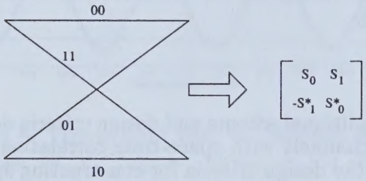

00
11
01
10
$\begin{bmatrix} S_0 & S_1 \\ -S^*_1 & S^*_0 \end{bmatrix}$

Figure 10: 0.5 bit/symbol 2-state BPSK smart-robust code

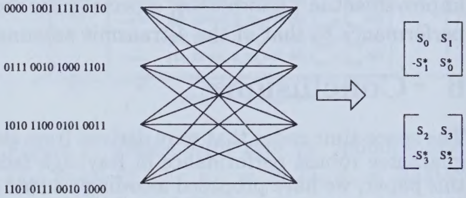

0000 1001 1111 0110
0111 0010 1000 1101
1010 1100 0101 0011
1101 0111 0010 1000
$\begin{bmatrix} S_0 & S_1 \\ -S^*_1 & S^*_0 \end{bmatrix}$
$\begin{bmatrix} S_2 & S_3 \\ -S^*_3 & S^*_2 \end{bmatrix}$

Figure 11: 4-state, 0.5 bit/symbol, BPSK smart-robust code for 2 transmit antennas

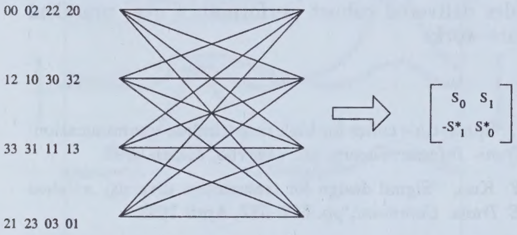

00 02 22 20
12 10 30 32
33 31 11 13
21 23 03 01
$\begin{bmatrix} S_0 & S_1 \\ -S^*_1 & S^*_0 \end{bmatrix}$

Figure 12: 1 bit/symbol QPSK smart-robust code

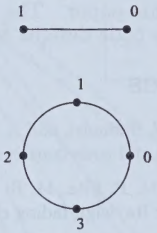


Figure 13: BPSK and QPSK constellations

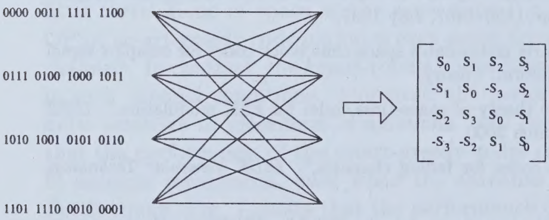

0000 0011 1111 1100
0111 0100 1000 1011
1010 1001 0101 0110
1101 1110 0010 0001
$\begin{bmatrix} S_0 & S_1 & S_2 & S_3 \\ -S_1 & S_0 & -S_3 & S_2 \\ -S_2 & S_3 & S_0 & -S_1 \\ -S_3 & -S_2 & S_1 & S_0 \end{bmatrix}$

Figure 14: 0.5 bit/symbol BPSK smart-robust code for 4 transmit antennas

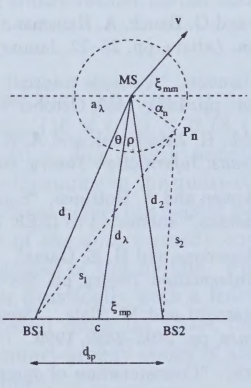

v
MS
ξ_{mm}
a_λ
α_n
θ
ρ
P_n
d_1
d_2
d_λ
s_2
s_1
ξ_{mp}
BS1
c
BS2
d_{sp}

Figure 15: The ring of scatterers model for multiple base station antennas

Signal Constellations for Noncoherent Space-Time Communications

Michael L. McCloud, Matthias Brehler, and Mahesh K. Varanasi
Department of Electrical and Computer Engineering
University of Colorado at Boulder
425 UCB Boulder, CO 80309 USA

Abstract

We examine the problem of designing signal constellations for the multiple transmit-multiple receive antenna Rayleigh fading communication channel. By employing the asymptotic union bound on the probability of error, we find a new formulation of the problem of signal design for the noncoherent fading channel. Since unitary signals are optimal (in the limit of large signal to noise ratios), the design can be posed in terms of packings on the Grassmanian manifold. A key difference in our approach from that of other authors is that we use a notion of distance on this manifold which is suggested by the union bound. As a consequence of our use of this distance measure we obtain signals designs that are guaranteed to achieve the full diversity order of the channel, a claim which does not hold when the chordal distance is used. We introduce a new method of iteratively designing signals, termed successive updates, to approximately optimize this performance measure.

1 Introduction

In this paper we consider the problem of designing signal waveforms for use with multiple transmit and receive antennas over the wireless fading channel, also known as space-time communication. This problem has received considerable attention, due in large part to the information theoretical work of Telatar, Foschini, and Gans [1, 2], who showed that the use of transmit (and receive) antenna diversity could achieve considerable gains on the Rayleigh fading channel when the receiver has perfect side information about the channel state. This work was later extended to the noncoherent channel where the receiver is not allowed this side information in [3]. In that paper the capacity achieving signal designs were shown to have a unitary structure. The asymptotic structure of the noncoherent channel capacity was further explored in [4], where the problem was related to sphere packing on the Grassmanian manifold. The common theme in these papers is that diversity communications can yield significant improvements over single antenna communications.

In this paper we consider the *asymptotic union bound* (AUB) on the probability of error for a noncoherent space-time communication system operating over a Rayleigh fading channel. The AUB relates the error probability to a relatively simple algebraic function of the signals

---

This work was supported in part by NSF Grants ANI-9725778 and ECS-9979400.

and hence can be used to generate a metric for designing space-time signal constellations. We introduce a method of iteratively designing signals, termed successive updates, to approximately optimize this performance measure. We also use the AUB to suggest modifications to the systematic designs of [5] and to the Grassmanian manifold based optimization of [6]. In particular, we are able to insure that the full potential order of diversity of the space-time channel is realized through our designs, a result which does not follow from the design metric employed in [5] and [6].

2 The Rayleigh Fading Channel

We adopt the usual block fading channel model for space-time communications wherein the discrete time received signal at the n^{th} antenna is of the form

$$\mathbf{y}_n = \sqrt{\overline{\gamma}}\mathbf{S}_i\mathbf{h}_n + \mathbf{n}_n. \tag{1}$$

Here $\mathbf{S}_i \in \mathbb{C}^{D\times M}$ is the i^{th} signal from a set of cardinality S, $\mathbf{h}_n \in \mathbb{C}^{M\times 1}$ is the vector of fading coefficients from each transmit antenna to sensor n, and $\mathbf{n}_n$ is additive noise.

Stacking the D observations per receiver antenna yields the DN sufficient statistics

$$\mathbf{y} = [\mathbf{y}_1^{\mathsf{T}}, \ldots, \mathbf{y}_N^{\mathsf{T}}]^{\mathsf{T}} = \sqrt{\overline{\gamma}}\boldsymbol{\mathcal{S}}_i\mathbf{h} + \mathbf{n}, \tag{2}$$

where $\boldsymbol{\mathcal{S}}_i = \mathbf{I}_N \otimes \mathbf{S}_i$ ($\otimes$ is the Kronecker product and $\mathbf{I}_N$ is the $N \times N$ identity matrix), $\mathbf{h} = [\mathbf{h}_1^{\mathsf{T}}, \ldots, \mathbf{h}_N^{\mathsf{T}}]^{\mathsf{T}}$, and $\mathbf{n} = [\mathbf{n}_1^{\mathsf{T}}, \ldots, \mathbf{n}_N^{\mathsf{T}}]^{\mathsf{T}}$. We assume that the additive noise is spatially white, but the fading need not be, i.e. $\boldsymbol{\Sigma} = E[\mathbf{h}\mathbf{h}^*]$ need not be diagonal. For the purposes of this paper we shall assume that "unitary" signals are used so that $\mathbf{S}_i^*\mathbf{S}_i = \mathbf{I}$. This assumption is justified by information theoretic arguments in [3] and by asymptotic error probability arguments in [7], each for the large SNR regime. Since our signal designs will be based on a distance measure valid for that regime, this assumption is not restrictive. In this model we normalize the noise so that $E[\mathbf{n}\mathbf{n}^*] = \mathbf{I}$, and the fading process so that $E[\mathbf{h}_n^*\mathbf{h}_n] = 1$. We denote the average SNR at each receiver antenna by the scalar $\overline{\gamma}$.

Conditioned on the transmission of signal $\boldsymbol{\mathcal{S}}_i$, the measurement has likelihood function

$$f_i(\mathbf{y}) = \frac{1}{\pi^{ND}\left|\mathbf{I}_{DN} + \overline{\gamma}\boldsymbol{\mathcal{S}}_i\Sigma\boldsymbol{\mathcal{S}}_i^*\right|} \exp\left\{-\mathbf{y}^*\left(\mathbf{I}_{DN} + \overline{\gamma}\boldsymbol{\mathcal{S}}_i\Sigma\boldsymbol{\mathcal{S}}_i^*\right)^{-1}\mathbf{y}\right\}. \tag{3}$$

The optimal detection strategy is the maximum likelihood rule [1]

$$\phi^{ML} : \widehat{i} = \arg\max_i f_i(\mathbf{y}) \;=\; \arg\max_i \mathbf{y}^*\boldsymbol{\mathcal{S}}_i\left(\mathbf{I}_{DN} + \overline{\gamma}^{-1}\Sigma^{-1}\right)^{-1}\boldsymbol{\mathcal{S}}_i^*\mathbf{y} \tag{4}$$

where we have applied the Woodbury identity to the quadratic form in the likelihood function.

3 Performance of the Maximum Likelihood Detector

To evaluate the probability of error for the maximum likelihood test of (4) we employ a union upper bound on the probability of error:

$$P^e \leq \frac{1}{S}\sum_{i=1}^{S}\sum_{\substack{k=1 \\ i\neq k}}^{S} P_{i,k}, \tag{5}$$

---

[1] We shall assume that the fading correlation is full rank. If $\boldsymbol{\Sigma}$ is singular with rank r, we may form the eigen-decomposition $\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^*$ with $\mathbf{U} \in \mathbb{C}^{NM\times r}$. We can then replace the signals by $\{\boldsymbol{\mathcal{S}}_i\mathbf{U}\}$ and treat the fading as full rank with correlation $\boldsymbol{\Lambda}$.

where P^e is the probability of a symbol error and $P_{i,k}$ is the probability that signal $\boldsymbol{\mathcal{S}}_k$ is preferred by the detector to signal $\boldsymbol{\mathcal{S}}_i$, when $\boldsymbol{\mathcal{S}}_i$ was transmitted.

Defining the vectors $\mathbf{v}_i = \left(\mathbf{I}_{DN} + \overline{\gamma}^{-1}\Sigma^{-1}\right)^{-1/2} \boldsymbol{\mathcal{S}}_i^* \mathbf{y}$ and $\mathbf{v}_k = \left(\mathbf{I}_{DN} + \overline{\gamma}^{-1}\Sigma^{-1}\right)^{-1/2} \boldsymbol{\mathcal{S}}_k^* \mathbf{y}$, where $(\mathbf{I} + \Sigma^{-1})^{1/2}$ is the (unique) positive definite square root of the matrix $\mathbf{I} + \Sigma^{-1}$, we make an error whenever $||\mathbf{v}_i||^2 < ||\mathbf{v}_k||^2$. If we define the matrix

$$\mathbf{J} = \begin{bmatrix} \mathbf{I} & 0 \\ 0 & -\mathbf{I} \end{bmatrix} \tag{6}$$

and the vector $\mathbf{v} = [\mathbf{v}_i^{\mathrm{T}} \; \mathbf{v}_k^{\mathrm{T}}]$, the pairwise error probability is found to be

$$P_{i,k} = P\left[\beta = \mathbf{v}^* \mathbf{J} \mathbf{v} < 0\right]. \tag{7}$$

The characteristic function for β can be found using the results of [8, Appendix B] and is given by

$$G_\beta(s) = \frac{1}{|\mathbf{I} + s\mathbf{R}_{\mathbf{vv}}\mathbf{J}|}, \tag{8}$$

where

$$\mathbf{R}_{\mathbf{vv}} \;=\; E[\mathbf{v}\mathbf{v}^*] = \begin{bmatrix} \overline{\gamma}\Sigma & \mathcal{F}\left(\mathbf{I} + \overline{\gamma}\Sigma\right)\boldsymbol{\mathcal{C}}_{i,k}\mathcal{F} \\ \mathcal{F}\boldsymbol{\mathcal{C}}_{i,k}^*\left(\mathbf{I} + \overline{\gamma}\Sigma\right)\mathcal{F} & \mathcal{F}\left(\mathbf{I} + \overline{\gamma}\boldsymbol{\mathcal{C}}_{i,k}^*\Sigma\boldsymbol{\mathcal{C}}_{i,k}\right)\mathcal{F} \end{bmatrix}$$

and $\mathbf{C}_{i,k} = \mathbf{S}_i^*\mathbf{S}_k$, $\boldsymbol{\mathcal{C}}_{i,k} = \boldsymbol{\mathcal{S}}_i^*\boldsymbol{\mathcal{S}}_k = \mathbf{I}_N \otimes \mathbf{C}_{i,k}$, and $\mathcal{F} = \left(\mathbf{I} + \overline{\gamma}^{-1}\Sigma^{-1}\right)^{-1/2}$. The performance of the detector depends on the non-zero eigenvalues, $\{\gamma_k\}_1^{2L}$, of the product $\mathbf{R}_{\mathbf{vv}}\mathbf{J}$ through

$$P_{i,k} = -\sum_{\gamma_j < 0} \operatorname{Res}\left(\frac{1}{s\prod_{l=1}^{2L}\gamma_l\left(s + \frac{1}{\gamma_l}\right)}, s_j = -\frac{1}{\gamma_j}\right), \tag{9}$$

summing over the residues of the negative eigenvalues (see e.g. [7]). Notice that this bound holds whether or not full diversity order is achieved. When full diversity order is achieved we have $L = MN$.

3.1 The Asymptotic Union Bound

We consider the behavior of the pairwise error probabilities with arbitrary Σ and M as the SNR approaches infinity. Denote the asymptotic $L \leq MN$ positive eigenvalues of the matrix $\mathbf{R}_{\mathbf{vv}}\mathbf{J}$ by $\alpha_l^{i,k}$, i.e. $\alpha_l^{i,k}/\overline{\gamma} = \lim_{\overline{\gamma}\to\infty} \gamma_l^{i,k}/\overline{\gamma}$. We find that the pairwise error probabilities approach [7]

$$P_{ik} \to \binom{2L}{L} \frac{1}{\prod_{l=1}^{L}\alpha_l^{i,k}}. \tag{10}$$

When full diversity is achieved we have $L = MN$ and the corresponding error probabilities have the form

$$P_{ik} \to \binom{2MN}{MN} \frac{\overline{\gamma}^{-MN}}{|\Sigma||\mathbf{I} - \mathbf{C}_{ik}\mathbf{C}_{ik}^*|^N}. \tag{11}$$

The corresponding union bound, denoted as the *asymptotic union bound* is given by

$$P^{AUB} = \frac{1}{S}\sum_{i=1}^{S}\sum_{k\neq i} P_{ik} = \frac{1}{S}\binom{2MN}{MN}\frac{\overline{\gamma}^{-MN}}{|\Sigma|}\sum_{i=1}^{S}\sum_{k\neq i}\frac{1}{|\mathbf{I} - \mathbf{C}_{ik}\mathbf{C}_{ik}^*|^N}. \tag{12}$$

Notice that in the large SNR regime, the dependence of the bound on the signal geometry (through the determinant $|\mathbf{I} - \mathbf{C}_{ik}\mathbf{C}_{ik}^*|^N$) is decoupled from the fading correlation structure. That means that we can design signals without paying attention to the particular correlation structure if we use design techniques based on this large SNR performance measure.

4 Signal Design for the Rayleigh Fading Channel

To develop an optimality criterion for the signal design problem, we need a numerical function which characterizes (asymptotically) the performance of the detector as a function of the signal constellation. As argued in the preceding section, the probability of error for the Rayleigh fading channel is asymptotically determined by the function

$$d(\{\mathbf{S}_i\}) \;=\; \sum_{i=1}^{S}\sum_{k\neq i} \frac{1}{|\mathbf{I} - \mathbf{S}_i^*\mathbf{S}_k\mathbf{S}_k^*\mathbf{S}_i|^N}. \tag{13}$$

Notice that this metric differs from the worst-case chordal distance employed in [5] and [6], which is defined by

$$d_c(\{\mathbf{S}_i\}) \;=\; \min_{k\neq i} \operatorname{tr}\left\{\mathbf{I} - \mathbf{S}_i^*\mathbf{S}_k\mathbf{S}_k^*\mathbf{S}_i\right\}. \tag{14}$$

Optimizing signal constellations over this latter metric does *not* guarantee that full potential diversity order of the space-time channel is achieved, a point which is discussed further in Section 6.

With the measure of distance given by (13), we may state the signal design problem as follows: Given $M, N, D, S \in \mathbb{N}$ ($\mathbb{N}$ is the set of natural numbers) find the set $\{\mathbf{S}_i\}_{i=1}^S$ for which the value of $d(\{\mathbf{S}_i\})$ in (13) is minimized subject to the constraints 1) $\mathbf{S}_i{}^*\mathbf{S}_i = \mathbf{I}$ for all i, 2) $\mathbf{S}_i \in \mathbb{C}^{D\times M}$.

Since we constrain our signals to have a unitary structure, their properties are completely determined by the M-dimensional subspace, $< \mathbf{S}_i >$, which they span. This means that we may parameterize our optimization over the Grassmanian manifold $G(M, D)$, the space of all M-dimensional subspaces of $\mathbb{C}^D$. A parameterization of $G(M, D)$ which employs $2MD - M^2$ real parameters is explicitly detailed in [6] and was employed for our designs. The AUB provides a smooth differentiable function of these parameters and standard gradient based optimization may be performed to design the signals. Unfortunately, as the signal set cardinality grows large (corresponding to large spectral efficiencies), this optimization becomes infeasible due to the number of free parameters. For this reason we consider a suboptimal iterative procedure for approximately optimizing this function in the next section.

5 Successive Updates for Signal Design

In order to solve the signal design problem of the previous section, an optimization over $S(2DM - M^2)$ parameters describing the signals must be performed. In order to find a tractable procedure for designing large cardinality signal sets, we consider a suboptimal approach, termed *successive updates*.

Suppose that we have already chosen K signals and seek an additional signal, $\mathbf{S}_{K+1}$ to add to the set such that we increase $d(\{\mathbf{S}_k\})$ by as little as possible. The AUB metric may be expanded as

$$d(\{\mathbf{S}_k\}_{k=1}^{K+1}) = d(\{\mathbf{S}_k\}_{k=1}^{K}) + 2\sum_{k=1}^{K} \frac{1}{|\mathbf{I} - \mathbf{S}_{K+1}^{*}\mathbf{S}_k\mathbf{S}_k^{*}\mathbf{S}_{K+1}|^{N}}. \tag{15}$$

The strategy we employed was to minimize the second term in (15) under the constraint that the additional signal has a unitary structure, $\mathbf{S}_{K+1}^{*}\mathbf{S}_{K+1} = \mathbf{I}$. We initialized each optimization step with a random "unitary" signal. The optimization problem can be parameterized on the Grassmanian manifold as in Section 4.

It is easy to extend this algorithm to add multiple signals at each iteration. In the limit we have the full optimization problem described in Section 4. We can also use the signals designed with this algorithm with $S' < S$ updates per iteration to initialize the full, non-iterative, optimization.

5.1 Numerical results for M=1

We first consider the case of a single transmit antenna, using the successive updates algorithm with one signal per update. In Figure 1 we plot the probability of symbol error versus the SNR for 4 receive antennas and $D = 2$ dimensional constellations for several values of the spectral efficiency. We notice that we incur a $6dB$ loss in SNR for every increase by one in the spectral efficiency. This is the behavior predicted in [4] for the capacity of this channel, i.e. that the spectral efficiency should increase by $M^{*}(1 - M^{*}/D)$bps/Hz for every $3dB$ gain in SNR, where $M^{*} = min\{M, N, \lfloor D/2 \rfloor\}$. These results imply that our signal designs display the relationship between spectral efficiency and SNR expected of optimal constellations.

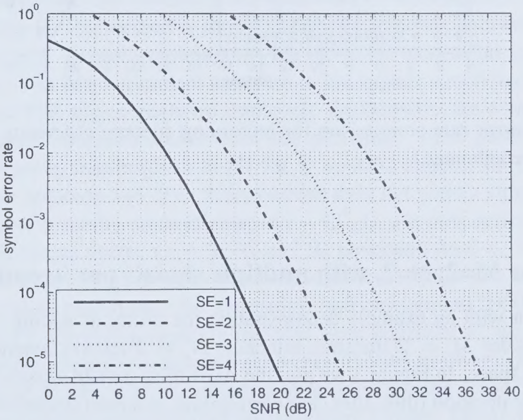


Figure 1: Probability of symbol error versus SNR for the $M = 1$ transmit, $N = 4$ receive antenna channel with $D = 2$ dimensional signal constellations for several values of the spectral efficiency.

5.2 Results for M=N=2

We now present results for multiple transmit and receive antennas. Using the successive updates algorithm with one signal added per update, a signal set with $S = 512$ signals in $D = 4$ dimensions for $M = N = 2$ transmit and receive antennas was designed. The spectral efficiency of the set is hence 2.25bps/Hz. To show the price in energy-efficiency for this relatively high spectral efficiency, the performance of the set using only the first 16, 64, 256 signals, respectively, is also shown in Figure 2. The figure shows simulated symbol error rates and the asymptotic union bounds on the symbol error probability. Note that the latter are a good indicator of the signals sets' performances. From the figure we see that the price paid for increasing the spectral efficiency from 1bps/Hz to 2bps/Hz is roughly 6dB, indicating a 3dB penalty when compared to the capacity results in [4]. This gap is most likely due to the poor fit between the linear approximation of [4] to the capacity curve at the spectral efficiencies considered.

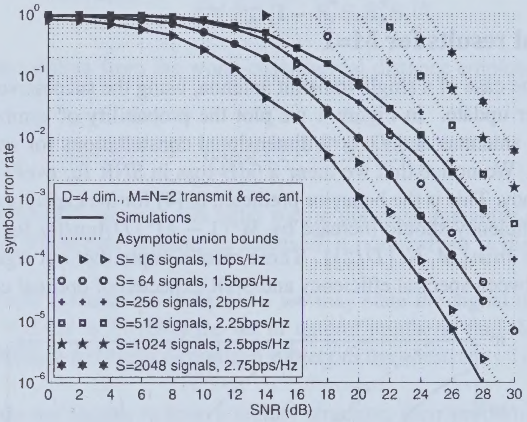


Figure 2: Symbol error rate comparison for increasing spectral efficiency for $M = N = 2$ transmit and receive antennas.

5.3 Results for M=2, N=2, with multiple signals per iteration

We end this section with an example demonstrating the utility of adding several signals per iteration. We consider $M = 2$ transmit antennas, $N = 2$ receive antennas and a $D = 4$ dimensional signal space. In Figure 3 we plot the AUB for $S = 128$ signals built with one and with 128 signals per iteration (this latter case corresponds to a joint optimization over all of the signals as in Section 4). We used the signals designed with one signal per iteration as an initial set for the joint optimization problem. We gained about 1dB in performance through the joint optimization relative to the signal by signal update for this example.

6 Modifications to Existing Techniques

The signal design techniques of [5] and [6] are based on the maximization of the worst-case chordal distance metric of (14). We can reformulate the algorithms presented in those refer-

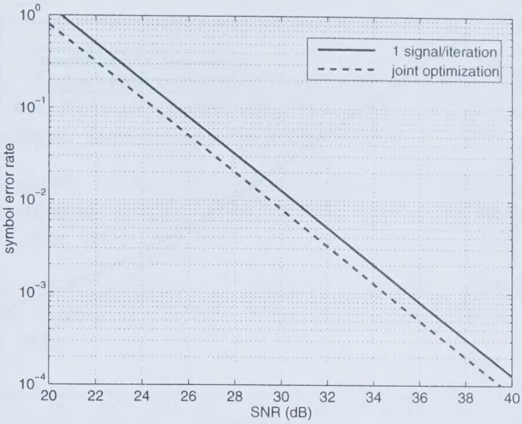


Figure 3: Symbol error rate comparison for signals designed with one signal update per iteration and with the joint optimization.

ences with the worst case chordal distance metric replaced by the AUB criterion of (13). The details of the resulting designs are presented in the subsequent sections.

6.1 Systematic Designs of Hochwald et. al

In [5] signal constellations are systematically generated by starting with an initial signal and successively rotating it. The method is shown to have an algebraic equivalence to block-codes in the ring of integers. Signal constellations are found by randomly searching the generating code matrices. The block-circulant structure in the signals simplifies the modulation and demodulation since PSK-type signals are employed. Furthermore, the design criterion of maximizing the minimum chordal distance does not necessarily yield good constellations. For example, the generating matrix given in [5] for $S = 2304$ signals, $M = 2$ transmit antennas, $N = 1$ receive antenna, and $D = 8$ dimensions does not yield a signal constellation that achieves full order of diversity, since the matrix $\mathbf{I} - \mathbf{C}_{i,k}^{*}\mathbf{C}_{i,k}$ is rank deficient for $\mathbf{S}_i$ generated from the (algebraic) input vector $\mathbf{x}_i = [0,\ 0]$ and $\mathbf{S}_k$ generated from $\mathbf{x}_k = [24,\ 24]$ (because of the cyclic structure of the code other input vectors yield similar low rank matrices). With a rather limited random search over the generating matrices (around 2000 matrices in a space of $48^{12} \approx 1.5 \cdot 10^{20}$ possible generating matrices were considered), we found a generating matrix[2] whose corresponding constellation achieves full order of diversity. Moreover, constellations with the same exponent in the union bound seem to be quite dense, meaning that several matrices were found which generated roughly equivalent signals. The asymptotic union bound on the probability of error of each design is shown in Figure 4. It is clear that at high SNRs the loss of diversity is significant, resulting in ever widening gap between the performance of the two designs. The minimum chordal distance of our constellation is 0.4062, while that of the constellation given in [5] is 0.4427, indicating the limited validity of the chordal distance as a figure of merit.

---

[2] $\begin{bmatrix} 3 & 26 & 14 & 10 & 2 & 46 \\ 1 & 30 & 39 & 47 & 38 & 16 \end{bmatrix}$.

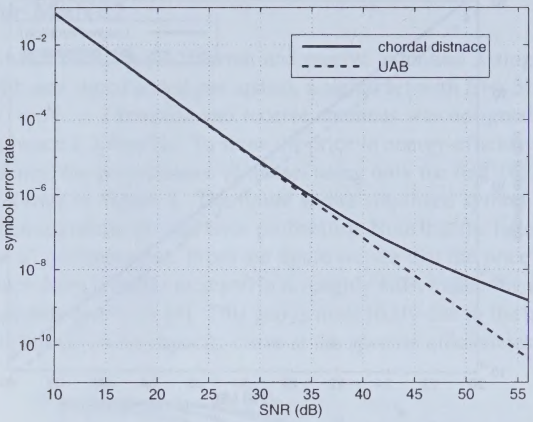


Figure 4: Probability of symbol error versus SNR for the $M = 2$ transmit, $N = 1$ receive, antenna channel for the $S = 2043$ constellation of [5], using the chordal distance on $G(M, D)$, and of this section, using the asymptotic union bound.

6.2 Sphere Packing on the Grassmanian Manifold

In [6], a sphere packing problem on $G(M, D)$, is described which is essentially the same as the joint optimization described in Section 4 with the worst case chordal distance of (14) employed in the place of (13). The authors of that paper used a relaxation technique to iteratively approximate the worst-case distance by smooth functions of the signal parameters.

A strong limitation to the sphere packing formulation of the signal design problem is the use of the minimum chordal distance as the objective function. This results in an l_∞ optimization problem which may not guarantee full diversity order. By using the asymptotic union bound of (12) we have an objective function which is a smooth differentiable function of the signal parameters and will guarantee that the full diversity order is achieved. We use a gradient search algorithm to solve the design problem, with gradients computed with respect to the Grassmanian parameters as in [6].

7 Conclusions

In this paper we have presented an approach to noncoherent signal design under dimensionality constraints for the multi-sensor Rayleigh fading channel based on the asymptotic union bound. By employing this distance measure, we can adapt the techniques of [5] and [6] to better reflect the error probability on the space time channel. We have also introduced a new technique, termed successive updates, which performs the signal design in an iterative fashion. The performance of our designs for two and three dimensions was shown to fit well with the theory of optimal designs presented in [4] for the case of single transmit antenna, while a $3dB$ gap was seen for the multiple transmit antenna case. This disparity is perhaps due to the failure of the linear approximation to the channel capacity at the spectral efficiencies considered and also appears in the designs of [5] and [6].

References

[1] İ. E. Telatar, "Capacity of multi-antenna gaussian channels," *European Trans. on Telecommun.*, vol. 10, no. 6, pp. 585–595, Nov. 1999, Originally a Bell Laboratories, Lucent Technologies, Technical Report, October 1995.

[2] G. J. Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Communications*, vol. 6, no. 3, pp. 311–335, Mar. 1998.

[3] T. L. Marzetta and B. M. Hochwald, "Capacity of a mobile multiple-antenna communication link in Rayleigh flat fading," *IEEE Trans. Inform. Theory*, vol. 45, no. 1, pp. 139–157, Jan. 1999.

[4] L. Zheng and D. N. C. Tse, "Packing spheres into the Grassmann manifold: A geometric approach to noncoherent multi-antenna channels," submitted to *IEEE Trans. Inform. Theory*, Apr. 2000.

[5] B. Hochwald, T. L. Marzetta, T. J. Richardson, W. Sweldens, and R. Urbanke, "Systematic design of unitary space-time constellations," *IEEE Trans. Inform. Theory*, vol. 46, no. 6, pp. 1962–1973, Sept. 2000.

[6] D. Agrawal, T. J. Richardson, and R. Urbanke, "Multiple-antenna signal constellations for fading channels," submitted to *IEEE Trans. Inform. Theory*, Nov. 1999.

[7] M. Brehler and M. K. Varanasi, "Asymptotic error probability analysis of quadratic receivers in Rayleigh fading channels with applications to a unified analysis of coherent and noncoherent space-time receivers," submitted to *IEEE Trans. Inform. Theory*, Aug. 2000.

[8] M. Schwartz, W. R. Bennett, and S. Stein, *Communication Systems and Techniques*, An IEEE Press Classic Reissue, New York, 1996, Originally A McGraw-Hill Publication, 1966.

[9] A. Björck and G. Golub, "Numerical methods for computing angles between linear subspaces," *Mathematics of Computation*, vol. 27, no. 123, pp. 579–594, July 1973.

[10] J. H. Conway, R. H. Hardin, and N. J. A. Sloane, "Packing lines, planes, etc: Packings in Grasssmannian spaces," *Experimental Mathematics*, vol. 5, pp. 139–59, 1996.

[11] G. Golub and C. F. V. Loan, *Matrix Computation*, Johns Hopkins Press, Baltimore, Baltimore, MD, 1991.

[12] B. M. Hochwald and T. L. Marzetta, "Unitary space-time modulation for multiple-antenna communications in Rayleigh flat fading," *IEEE Trans. Inform. Theory*, vol. 46, no. 2, pp. 543–564, Mar. 2000.

Space-Time Coded Generalized Multicarrier CDMA with Block-Spreading for Multirate Services*

Anastasios Stamoulis, Zhiqiang Liu, and Georgios B. Giannakis
Dept. of Electrical and Computer Engineering
University of Minnesota, Minneapolis, MN 55455
`{stamouli,lzq,georgios}@ece.umn.edu`

Abstract

Relying on space-time generalized multicarrier CDMA and exploiting both transmit and receive antenna diversity, we design herein multirate transceivers which guarantee deterministic symbol recovery with multipath-diversity gains regardless of the (possibly unknown) frequency-selective FIR channels and multiuser interference. Our approach is based on a three-level design of user codes: the outer code handles multiuser interference, the middle code results in space-time diversity gains, and the inner code mitigates intersymbol interference. In a multiuser/multirate setting, the use of space-time block coding with two transmit and a single receive antenna achieves guaranteed diversity gains, whereas the use of two receive antennas could potentially double the capacity of the system (in terms of maximum number of users or achievable transmission rates). Simulations illustrate the merits of our designs.

1 Introduction

Broadband wireless networks are envisioned to provide high data rates and support Integrated Services. To make them possible, the physical layer should i) cope with the variable channel capacity due to multipath fading; and ii) facilitate multiuser/multirate transmissions. Time, frequency, and spatial diversity have been employed to combat channel fading and the high bit-error-rates (BERs) it induces (see e.g., [6, 7] and references therein). Among the three, spatial diversity especially with transmit-antennas has drawn a lot of attention recently as it promises to increase the achievable transmission rates significantly.

The merits of block-spreading with an inner/outer code design have been described in [12] for block transmission systems. On the other hand, ST coding capitalizes on multiple transmit antennas to increase considerably the channel capacity [4]. Though it is well known that array processing and receive antennae improve BER performance, the size–power limitations of handheld devices make the deployment of multiple transmit antennas at the basestation an attractive alternative. ST coding can achieve spatial diversity gains with just a single receive antenna; as we show in this work, if the receiver is equipped with more than one receive antenna, further performance improvement can be achieved.

However, in broadband wireless networks, ST coding must take into consideration channel frequency-selectivity [6]. ST decoding schemes based on maximum-likelihood estimators may not be attractive, given that computational complexity grows exponentially with the transmission rate. Decoupling ST decoding from channel equalization appears to be a computationally

---

*Work in this paper was supported by the NSF Wireless Initiative grant no. 99-79443.

attractive alternative, and it has already been proposed in the form of ST coded OFDM [1]. However, the inner/outer code designs of [12] exhibit improved BER performance with respect to OFDM, which in part motivated the development of *single-rate* ST block codes in [5] with *a single receive-antenna*, and multi-rate ST block codes in [11].

In this paper we develop a novel multirate/multiuser scheme relying on space-time (ST) coded Generalized MultiCarrier (GMC) CDMA that exploits both transmit and receive antenna diversity. The resulting transceivers rely on block-spreading to guarantee symbol recovery with multipath-diversity gains regardless of the (possibly unknown) frequency-selective FIR channels and multiuser interference (MUI). Our approach is based on a three-level user code design: the outer code handles MUI, the middle code results in ST diversity gains, and the inner code mitigates inter-symbol interference (ISI).

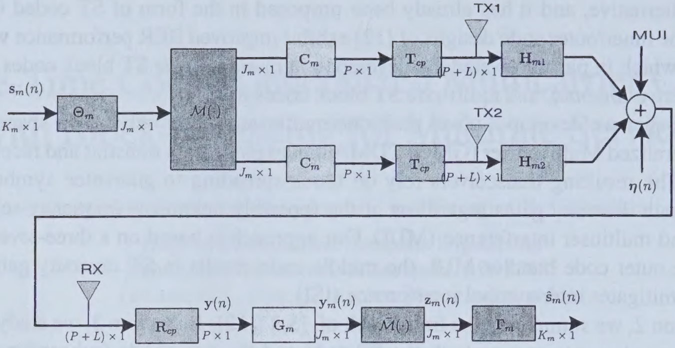
In Section 2, we summarize the framework of [5,11,12]. In Section 3, we study the effects of multiple receive antennas on both the ST code and the outer-code (subcarrier allocation). Having in mind that receive antenna diversity can be used to separate users in space, we show that, while ensuring ST diversity gains, either transmission rates or the maximum number of active users can be doubled just by deploying a second receive antenna. The merits of our system are verified through simulations in Section 4. Finally, Section 5 concludes and gives pointers to future research.

2 Physical Layer

First, we provide a high level view of our physical layer and then we provide a detailed mathematical description.

High-level View: At the physical layer there are $N_t = 2$ transmit-antennas and $N_r = 1$ receive-antenna for each of M users, as shown in Fig. 1. User m, $m = 0, \ldots, M-1$, relies on a block-spreading code to transmit information symbol blocks $\mathbf{s}_m(n)$ of size $K_m \times 1$. Through the redundant $J_m \times K_m$ precoder Θ_m, the blocks $\mathbf{s}_m(n)$ are mapped to $J_m \times 1$ blocks $\tilde{\mathbf{s}}_m(n)$; the ST mapper $\mathcal{M}(\cdot)$ maps two consecutive blocks $\tilde{\mathbf{s}}_m(2n)$, $\tilde{\mathbf{s}}_m(2n+1)$ to four blocks $\bar{\mathbf{s}}_{mi}(2n)$, $\bar{\mathbf{s}}_{mi}(2n+1)$, $i = 1, 2$ (where the index i denotes the transmit antenna): the coded blocks $\bar{\mathbf{s}}_{m1}(2n)$, $\bar{\mathbf{s}}_{m1}(2n+1)$ are to be parallel-to-serial converted, pulse-shaped and transmitted by the first antenna, and likewise $\bar{\mathbf{s}}_{m2}(2n)$, $\bar{\mathbf{s}}_{m2}(2n+1)$ are to be transmitted by the second antenna. The user-specific GMC-CDMA code matrix $\mathbf{C}_m$ maps each block to a $P \times 1$ transmitted block $\mathbf{u}_{mi}(n)$. Before the transmission of $\mathbf{u}_{mi}(n)$, the last L elements of $\mathbf{u}_{mi}(n)$ are copied to its beginning to avoid inter-block interference (IBI) caused by the FIR channel. This operation is represented by the matrix $\mathbf{T}_{cp}$ and implements the cyclic prefix (CP) as in OFDM. The $(P+L)$-long transmitted block goes through the FIR channel, which is denoted by the Toeplitz matrix $\mathbf{H}_{mi}$. As a result, the received block $\mathbf{y}(n)$ obtained after sampling the receive-filter output at the chip rate and discarding the CP, still contains MUI and additive noise $\eta(n)$ (see Fig. 1). The retrieval of the transmitted data $\mathbf{s}_m(n)$ (of the m-th user of interest) entails the following steps at the receiver: i) elimination of MUI which is achieved by the receive filterbank $\mathbf{G}_m$; ii) linear processing of the MUI-free block $\mathbf{y}_m(n)$ by the ST decoder $\bar{\mathcal{M}}(\cdot)$ to achieve transmit diversity gain regardless of the (possibly unknown) FIR channels; iii) elimination of ISI using the filterbank Γ_m. We will show later on how recovery of $\mathbf{s}_m(i)$ can be guaranteed regardless of the FIR channels.

Detailed View: Our model represents signals, codes and channels by chip-rate samples of their complex envelopes. The information symbols $s_m(k)$ of user m are grouped together in blocks $\mathbf{s}_m(n) {:=} \left(s_m(nK) \cdots s_m(nK+K-1) \right)^T$ of size K_m, which are mapped to blocks $\tilde{\mathbf{s}}_m(n) = \Theta_m \mathbf{s}_m(n)$ of length J_m through the tall $J_m \times K_m$ matrix Θ_m (the redundancy introduced



TX1
MUI
$\mathbf{s}_m(n)$
$K_m \times 1$
Θ_m
$J_m \times 1$
$\mathcal{M}(\cdot)$
$J_m \times 1$
$\mathbf{C}_m$
$P \times 1$
$\mathbf{T}_{cp}$
$(P+L) \times 1$
$\mathbf{H}_{m1}$
TX2
$J_m \times 1$
$\mathbf{C}_m$
$P \times 1$
$\mathbf{T}_{cp}$
$(P+L) \times 1$
$\mathbf{H}_{m2}$
+
$\eta(n)$
RX
$(P+L) \times 1$
$\mathbf{R}_{cp}$
$\mathbf{y}(n)$
$P \times 1$
$\mathbf{G}_m$
$\mathbf{y}_m(n)$
$J_m \times 1$
$\bar{\mathcal{M}}(\cdot)$
$\mathbf{z}_m(n)$
$J_m \times 1$
Γ_m
$\hat{\mathbf{s}}_m(n)$
$K_m \times 1$

Figure 1: Physical layer

by Θ_m facilitates ISI elimination and symbol recovery regardless of the physical channel [5, 12]). Then, the blocks $\tilde{\mathbf{s}}_m(n)$ are fed into the ST mapper $\mathcal{M}(\cdot)$, which takes two consecutive precoded blocks, $\tilde{\mathbf{s}}_m(2n)$ and $\tilde{\mathbf{s}}_m(2n+1)$, to output the following $2J_m \times 2$ code matrix ($*$ denotes conjugation):

$$\mathbf{S}(n) := \begin{pmatrix} \bar{\mathbf{s}}_{m1}(2n) & \bar{\mathbf{s}}_{m1}(2n+1) \\ \bar{\mathbf{s}}_{m2}(2n) & \bar{\mathbf{s}}_{m2}(2n+1) \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{s}}_m(2n) & -\tilde{\mathbf{s}}_m^*(2n+1) \\ \tilde{\mathbf{s}}_m(2n+1) & \tilde{\mathbf{s}}_m^*(2n) \end{pmatrix} \begin{array}{l} \rightarrow \quad \text{time} \\ \downarrow \quad \text{space} \end{array} \tag{1}$$

Each column corresponds to a transmission slot ("time"), whereas each row holds the symbols to be transmitted by each antenna ("space"). Note that without blocking (i.e., $J_m = 1$) the code matrix in (1) reduces to the ST block code in [2]. For notational convenience, we denote by $\bar{\mathbf{s}}_{mi}(n),\ i = 1, 2$, the $J_m \times 1$ block transmitted through the ith transmit-antenna at the time slot n, and observe from (1) that e.g., $\bar{\mathbf{s}}_{m1}(2n) = \tilde{\mathbf{s}}_m(2n)$, and $\bar{\mathbf{s}}_{m2}(2n+1) = \tilde{\mathbf{s}}_m^*(2n)$.

Instrumental in MUI elimination is the user-specific block-spreading code, which is represented by the $P \times J_m$ matrix $\mathbf{C}_m$. The user code is applied to $\bar{\mathbf{s}}_{mi}(n)$ to produce the block $\mathbf{u}_{mi}(n) := \mathbf{C}_m \bar{\mathbf{s}}_{mi}(n)$ of length P, which is transmitted by the m-th user's i-th antenna through the L-th order FIR channel $\{h_{mi}(l)\}_{l=0}^{L}$. In order to eliminate the IBI caused by FIR channels, similar to OFDM, we rely on inserting a CP of length L at the beginning of $\mathbf{u}_{mi}(n)$; at the receiver, the corresponding first L received symbols are discarded. As explained in [12] and shown in Fig. 1, the CP insertion can be described by the transmit-matrix $\mathbf{T}_{cp} := \left(\mathbf{I}_{cp}^T \ \ \mathbf{I}_P^T\right)^T$, where $\mathbf{I}_{cp}$ is formed by the last L rows of the $P \times P$ identity matrix $\mathbf{I}_P$. Correspondingly, the receiver operation of discarding the first L received symbols is described by the receive-matrix $\mathbf{R}_{cp} := \left(\mathbf{0}_{P \times L} \ \ \mathbf{I}_P\right)$. The effect of the FIR channel $h_{mi}(l)$ on the transmitted data is described by the $(P+L) \times (P+L)$ Toeplitz (convolutional) matrix $\mathbf{H}_{mi}$ with first row $\left(h_{mi}(0)\, \mathbf{0}_{1 \times (P+L-1)}\right)$, and first column $\left(h_{mi}(0) \ldots h_{mi}(L)\, \mathbf{0}_{1 \times (P-1)}\right)^T$. With $\tilde{\mathbf{H}}_{mi} := \mathbf{R}_{cp}\mathbf{H}_{mi}\mathbf{T}_{cp}$, the $P \times 1$ IBI-free received symbol block $\mathbf{y}(n)$ is given by (see Fig. 1):

$$\mathbf{y}(n) = \mathbf{x}_m(n) + \sum_{\mu=0,\ \mu \neq m}^{M-1} \mathbf{x}_\mu(n) + \mathbf{R}_{cp}\boldsymbol{\eta}(n)\ , \tag{2}$$

where: $\mathbf{x}_m(n) := \tilde{\mathbf{H}}_{m1}\mathbf{C}_m\bar{\mathbf{s}}_{m1}(n) + \tilde{\mathbf{H}}_{m2}\mathbf{C}_m\bar{\mathbf{s}}_{m2}(n)$ denotes the received symbol block from the two transmit-antennas of the mth user, and $\boldsymbol{\eta}(n)$ is the $(P+L) \times 1$ additive white Gaussian noise vector.

The received symbol block in (2) contains three terms: the first term is the useful received block $\mathbf{x}_m(n)$ from the desired mth user; the second term constitutes the MUI caused by simultaneous transmissions of the other $M-1$ users; the third term is the additive noise.

2.1 Code Design

In the following, we summarize our code design procedure starting from the outside ("outer-code") and moving to the inside ("inner-code") (further details can be found in [11]).
Outer Code: MUI Elimination The design of the outer code is based on the concept of subcarrier allocation. Similar to OFDMA, every user m is allocated a distinct set of J_m subcarriers to transmit information, but unlike OFDMA, the all-digital framework of [12] guarantees symbol recovery irrespective of channel nulls. Herein, we allocate J_m subcarriers to every user m; these subcarriers are used by both transmit antennas to produce two interfering signals at the receiver end (something that [12] did not have to cope with).

Each user μ is allocated a set $\mathcal{I}_\mu = \{j_{m,0}, \dots, j_{m,J_m-1}\}$ of J_μ subcarriers $\{\rho_{m,j_m}\}_{j_m=0}^{J_m-1}$; in our discrete-time equivalent baseband model, the j_m-th subcarrier is represented as $\rho_{m,j_m} = \exp\{j2\pi j_m/P\}$. Users have non-intersecting subcarrier sets, and the set of subcarriers is used to define for the μth user the subcarrier selector matrix $\mathbf{\Phi}_\mu = \mathbf{I}_P(:, \mathcal{I}_\mu)$ where $\mathbf{I}_P(:, \mathcal{I}_\mu)$ is a $P \times J_\mu$ matrix built from the $\mathcal{I}_\mu$ columns of $\mathbf{I}_P$. Denoting by $\mathbf{F}$ the $P \times P$ FFT matrix $\mathbf{F}$ with its (k,p) entry $[\mathbf{F}]_{k,p} = P^{-\frac{1}{2}} \exp\{(-j2\pi kp)/P\}$, the subcarrier selector $\mathbf{\Phi}_\mu$ is used to design the user coder $\mathbf{C}_\mu$ and the receive-filter $\mathbf{G}_\mu$ as:

$$\mathbf{C}_\mu = \mathbf{F}^{\mathcal{H}} \mathbf{\Phi}_\mu, \quad \mathbf{G}_\mu = \mathbf{\Phi}_\mu^T \mathbf{F}, \quad \mu = 0, \dots, M-1\,. \tag{3}$$

Then, it can be shown that the receive-filter output $\mathbf{y}_m(n) := \mathbf{G}_m \mathbf{y}(n)$ can be written as:

$$\mathbf{y}_m(n) = \boldsymbol{\mathcal{D}}_{m1} \bar{\mathbf{s}}_{m1}(n) + \boldsymbol{\mathcal{D}}_{m2} \bar{\mathbf{s}}_{m2}(n) + \mathbf{G}_m \mathbf{R}_{cp} \boldsymbol{\eta}(n), \tag{4}$$

where $\boldsymbol{\mathcal{D}}_{\mu i}$ is a $J_\mu \times J_\mu$ diagonal matrix holding in its diagonal the frequency responses $H_{mi}(\rho_{\mu,j_\mu}) = \sum_{l=0}^{L} h_{\mu i}(l) \rho_{\mu,j_\mu}^{-l}$ of the FIR channel $h_{\mu i}(l)$ at the corresponding subcarriers ρ_{μ,j_μ}'s. It is clear from (4) that the MUI has been eliminated through the joint designs of $\mathbf{C}_\mu$ and $\mathbf{G}_\mu$ in (3). Note also that the MUI elimination is achieved without requiring the knowledge of the channel impulse response (CIR) at the receiver.
Middle-Code: Diversity Gain The objective of the middle-code is to collect diversity gains by combining constructively the transmitted signals from the two antennas. We consider two consecutive MUI-free blocks: $\mathbf{y}_m(2n)$ and $\mathbf{y}_m(2n+1)$, which can be expressed as:

$$\mathbf{y}_m(2n) = \boldsymbol{\mathcal{D}}_{m1} \tilde{\mathbf{s}}_m(2n) + \boldsymbol{\mathcal{D}}_{m2} \tilde{\mathbf{s}}_m(2n+1) + \mathbf{G}_m \mathbf{R}_{cp} \boldsymbol{\eta}(2n) \tag{5a}$$

$$\mathbf{y}_m(2n+1) = -\boldsymbol{\mathcal{D}}_{m1} \tilde{\mathbf{s}}_m^*(2n+1) + \boldsymbol{\mathcal{D}}_{m2} \tilde{\mathbf{s}}_m^*(2n) + \mathbf{G}_m \mathbf{R}_{cp} \boldsymbol{\eta}(2n+1). \tag{5b}$$

In order to exploit the embedded diversity of (5a) and (5b), we design the ST decoder $\bar{\mathcal{M}}(\cdot)$ (see Fig. 1) by forming its two consecutive output blocks, $\mathbf{z}_m(2n)$ and $\mathbf{z}_m(2n+1)$, as:

$$\begin{pmatrix} \mathbf{z}_m(2n) \\ \mathbf{z}_m(2n+1) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mathcal{D}}_{m1}^* & \boldsymbol{\mathcal{D}}_{m2} \\ \boldsymbol{\mathcal{D}}_{m2}^* & -\boldsymbol{\mathcal{D}}_{m1} \end{pmatrix} \begin{pmatrix} \mathbf{y}_m(2n) \\ \mathbf{y}_m^*(2n+1) \end{pmatrix}, \tag{6}$$

and we arrive at: $\mathbf{z}_m(n) = [\boldsymbol{\mathcal{D}}_{m1}^* \boldsymbol{\mathcal{D}}_{m1} + \boldsymbol{\mathcal{D}}_{m2}^* \boldsymbol{\mathcal{D}}_{m2}] \tilde{\mathbf{s}}_m(n) + \tilde{\boldsymbol{\eta}}_m(n)$, where
$\tilde{\boldsymbol{\eta}}_m(n) := [\boldsymbol{\eta}_m^T(2n) \;\; \boldsymbol{\eta}_m^T(2n+1)]^T$. If the receiver noise is temporally white, then it can be seen that $\boldsymbol{\eta}(n)$, $\boldsymbol{\eta}_m(n)$ are white Gaussian random process with diagonal covariance matrices. As a

result, the blocks $\mathbf{z}_m(n)$ can be processed individually without loss in performance. Furthermore, we obtain for $\mathbf{z}_m(n)$:

$$\mathbf{z}_m(n) = \boldsymbol{\mathcal{D}}_m \bar{\mathbf{s}}_m(n) + \boldsymbol{\eta}_m(n) = \boldsymbol{\mathcal{D}}_m \boldsymbol{\Theta}_m \mathbf{s}_m(n) + \boldsymbol{\eta}_m(n), \tag{7}$$

where $\boldsymbol{\mathcal{D}}_m := \boldsymbol{\mathcal{D}}^*_{m1}\boldsymbol{\mathcal{D}}_{m1} + \boldsymbol{\mathcal{D}}^*_{m2}\boldsymbol{\mathcal{D}}_{m2} = \operatorname{diag}[\sum_{i=1}^2 |H_{mi}(\rho_{m,1})|^2, \ldots, \sum_{i=1}^2 |H_{mi}(\rho_{m,J_m})|^2]$.

As each channel can have at most L nulls, (7) implies that transmit-diversity gains are achieved in at least $J_m - 2L$ subcarriers. Moreover, $J_m - L$ subcarriers are guaranteed to be non-zero and can be used for symbol recovery (regardless of the underlying channels). The latter observation brings us to the third component of our (pre)coding, namely, the inner code.

Inner-Code: Guaranteed Symbol Recovery We deduce from (7) that the recovery of $\mathbf{s}_m(n)$ from $\mathbf{z}_m(n)$ requires $\boldsymbol{\mathcal{D}}_m\boldsymbol{\Theta}_m$ to be full column rank. As shown in [5, 12], the full column rank is guaranteed by selecting: i) $J_m = K_m + L$, and ii) any K_m rows of $\boldsymbol{\Theta}_m$ to be linearly independent. To satisfy conditions i) and ii), a special choice of $\boldsymbol{\Theta}_m$ is given by [12]:

$$\boldsymbol{\Theta}_m = \begin{pmatrix} 1 & \rho_{m,0}^{-1} & \cdots & \rho_{m,0}^{-(K_m-1)} \\ \vdots & & & \vdots \\ 1 & \rho_{m,J_m-1}^{-1} & \cdots & \rho_{m,J_m-1}^{-(K_m-1)} \end{pmatrix}. \tag{8}$$

3 Space-Time CDMA With Two Receive Antennas

In a single user environment, it is well known that more reliable symbol estimates can be obtained using more than one receive antenna. If we recall that in our system the outer code (subcarrier allocation) converts the multiuser environment to a single-user one, then single-user array-processing algorithms could be applied to our framework (with a few modifications) to improve the BER, if the receiver is equipped with $N_r > 1$ antennas. As we demonstrate in the simulations section, this is indeed the case in our framework. However, in a multiuser environment, multiple receive antennas can also be used for user separation; for example, M users can be separated at the receiver using M receive antennas. Hence, the presence of multiple receive antennas could be taken into account in the design of the outer code.

The key idea we pursue is that two users μ and m with same transmission rates ($J_m = J_\mu = J$) may share the same set of subcarriers and be separated using array processing. In more detail, when users m and μ have been allocated the same subcarriers (i.e., the same *outer code*), then after eliminating the other users, the received signal in the first antenna is (cf. (4)):

$$\mathbf{y}_{m,\mu}(n) = \underbrace{\boldsymbol{\mathcal{D}}_{m1}\bar{\mathbf{s}}_{m1}(n) + \boldsymbol{\mathcal{D}}_{m2}\bar{\mathbf{s}}_{m2}(n)}_{\text{user } m} + \underbrace{\boldsymbol{\mathcal{D}}_{\mu1}\bar{\mathbf{s}}_{\mu1}(n) + \boldsymbol{\mathcal{D}}_{\mu2}\bar{\mathbf{s}}_{\mu2}(n)}_{\text{user } \mu} + \mathbf{G}_{m,\mu}\mathbf{R}_{cp}\boldsymbol{\eta}(n)\,. \tag{9}$$

Though our designs can be generalized for $N_r > 2$ antennas, let us focus in the case of two receive antennas. Using the prime (′) to denote channels and received signals corresponding to the second receive antenna, we obtain:

$$\mathbf{y}'_{m,\mu}(n) = \underbrace{\boldsymbol{\mathcal{D}}'_{m1}\bar{\mathbf{s}}_{m1}(n) + \boldsymbol{\mathcal{D}}'_{m2}\bar{\mathbf{s}}_{m2}(n)}_{\text{user } m} + \underbrace{\boldsymbol{\mathcal{D}}'_{\mu1}\bar{\mathbf{s}}_{\mu1}(n) + \boldsymbol{\mathcal{D}}'_{\mu2}\bar{\mathbf{s}}_{\mu2}(n)}_{\text{user } \mu} + \mathbf{G}_{m,\mu}\mathbf{R}_{cp}\boldsymbol{\eta}'(n)\,. \tag{10}$$

By taking into account the ST mapping of (1) we obtain (cf. (5a) and (5b)):

$$\underbrace{\begin{pmatrix} \mathbf{y}_{m,\mu}(2n) \\ \mathbf{y}^*_{m,\mu}(2n+1) \\ \mathbf{y}'_{m,\mu}(2n) \\ \mathbf{y}'^*_{m,\mu}(2n+1) \end{pmatrix}}_{\mathbf{y}} = \underbrace{\begin{pmatrix} \begin{pmatrix} \boldsymbol{\mathcal{D}}_{m1} & \boldsymbol{\mathcal{D}}_{m2} \\ \boldsymbol{\mathcal{D}}^*_{m2} & -\boldsymbol{\mathcal{D}}^*_{m1} \end{pmatrix} & \begin{pmatrix} \boldsymbol{\mathcal{D}}_{\mu1} & \boldsymbol{\mathcal{D}}_{\mu2} \\ \boldsymbol{\mathcal{D}}^*_{\mu2} & -\boldsymbol{\mathcal{D}}^*_{\mu1} \end{pmatrix} \\ \begin{pmatrix} \boldsymbol{\mathcal{D}}'_{m1} & \boldsymbol{\mathcal{D}}'_{m2} \\ \boldsymbol{\mathcal{D}}'^*_{m2} & -\boldsymbol{\mathcal{D}}'^*_{m1} \end{pmatrix} & \begin{pmatrix} \boldsymbol{\mathcal{D}}'_{\mu1} & \boldsymbol{\mathcal{D}}'_{\mu2} \\ \boldsymbol{\mathcal{D}}'^*_{\mu2} & -\boldsymbol{\mathcal{D}}'^*_{\mu1} \end{pmatrix} \end{pmatrix}}_{\mathbf{H}} \underbrace{\begin{pmatrix} \bar{\mathbf{s}}_m(2n) \\ \bar{\mathbf{s}}_m(2n+1) \\ \bar{\mathbf{s}}_\mu(2n) \\ \bar{\mathbf{s}}_\mu(2n+1) \end{pmatrix}}_{\bar{\mathbf{s}}} + \underbrace{\begin{pmatrix} \mathbf{G}_{m,\mu}\mathbf{R}_{cp}\boldsymbol{\eta}(2n) \\ \mathbf{G}_{m,\mu}\mathbf{R}_{cp}\boldsymbol{\eta}(2n+1) \\ \mathbf{G}_{m,\mu}\mathbf{R}_{cp}\boldsymbol{\eta}'(2n) \\ \mathbf{G}_{m,\mu}\mathbf{R}_{cp}\boldsymbol{\eta}'(2n+1) \end{pmatrix}}_{\mathbf{n}} \tag{11}$$

which, for compactness, we write as:

$$\boldsymbol{y} = \boldsymbol{H}\tilde{\boldsymbol{s}} + \boldsymbol{n}\,, \tag{12}$$

where the vectors $\boldsymbol{y}$, $\tilde{\boldsymbol{s}}$, $\boldsymbol{n}$ are $4J \times 1$, and the matrix $\boldsymbol{H}$ is $4J \times 4J$. Note also that the diagonal "$\mathcal{D}$" matrices in (9) and (10) make each of the $2J \times 2J$ submatrices $\boldsymbol{H}_{1,1}, \boldsymbol{H}_{1,2}, \boldsymbol{H}_{2,1}, \boldsymbol{H}_{2,2}$ of

$$\boldsymbol{H} = \begin{pmatrix} \boldsymbol{H}_{1,1} & \boldsymbol{H}_{1,2} \\ \boldsymbol{H}_{2,1} & \boldsymbol{H}_{2,2} \end{pmatrix}$$

have the special structure shown in (11). It is naturally expected that users m, μ will also share the same inner code, denoted by $\Theta_{m,\mu}$. Therefore, (12) can also be written as:

$$\boldsymbol{y} = \tilde{\boldsymbol{H}}\boldsymbol{s} + \boldsymbol{n}\,, \tag{13}$$

where $\tilde{\boldsymbol{H}}{:=}\boldsymbol{H}\ (\mathrm{I}_{4J} \otimes \Theta_{m,\mu})$, and $\boldsymbol{s}{:=}\left(\ \mathrm{s}_m^T(2n)\ \mathrm{s}_m^T(2n+1)\ \mathrm{s}_\mu^T(2n)\ \mathrm{s}_\mu^T(2n+1)\ \right)^T$. Let us see now how the transmitted symbols can be recovered from (12) (or (13)).

3.1 Linear and DF receivers

If the receive antennas are well separated, the underlying physical channels are independent and the matrices $\boldsymbol{H}$, $\tilde{\boldsymbol{H}}$ are full column rank. Then, a number of techniques can be applied to (12) for recovering the transmitted symbols. For example, [8] offers closed form ZF and MMSE receiver designs, whereas [9] suggests a block decision feedback (DF) approach for improved BER. According to [8], the ZF receiver can be implemented by the filterbank $\boldsymbol{\Gamma}_{zf} = \boldsymbol{H}^\dagger$ (the pseudoinverse of $\boldsymbol{H}$), whereas the MMSE receiver is given by $\boldsymbol{\Gamma}_{mmse} = \mathrm{R}_{\tilde{s}\tilde{s}}\boldsymbol{H}^{\mathcal{H}}(\mathrm{R}_{nn} + \boldsymbol{H}\mathrm{R}_{\tilde{s}\tilde{s}}\boldsymbol{H}^{\mathcal{H}})^{-1}$, with $\mathrm{R}_{\tilde{s}\tilde{s}}{:=}E\{\tilde{\boldsymbol{s}}\tilde{\boldsymbol{s}}^{\mathcal{H}}\}$, $\mathrm{R}_{nn}{:=}E\{\boldsymbol{n}\boldsymbol{n}^{\mathcal{H}}\}$ the autocorrelation of $\tilde{\boldsymbol{s}}$ and $\boldsymbol{n}$, respectively. On the other hand, the DF designs of [9] are based on two filterbanks described by the matrices W, B which implement the feedforward and the feedback part of the DF receiver. The ZF-DF receiver $(\mathrm{W}_{zf}, \mathrm{B}_{zf})$ is given by: $\mathrm{W}_{zf} = \mathrm{D}_{zf}^{-1}\mathrm{U}_{zf}^{-\mathcal{H}}\tilde{\boldsymbol{H}}^{\mathcal{H}}\mathrm{R}_{nn}^{-1}$, $\mathrm{B}_{zf} = \mathrm{U} - \mathrm{I}_{4J}$, where U_{zf} is an upper triangular matrix with unit diagonal given by the Cholesky factorization of the matrix $\tilde{\boldsymbol{H}}^{\mathcal{H}}\mathrm{R}_{nn}^{-1}\tilde{\boldsymbol{H}} = \mathrm{U}_{zf}^{\mathcal{H}}\mathrm{D}_{zf}\mathrm{U}_{zf}$ and D_{zf} a diagonal matrix. The MMSE-DF $(\mathrm{W}_{mmse}, \mathrm{B}_{mmse})$ receiver takes into account both $\mathrm{R}_{sy}{:=}E\{\boldsymbol{s}\boldsymbol{y}^{\mathcal{H}}\}$ and $\mathrm{R}_{yy} = \mathrm{R}_{nn} + \tilde{\boldsymbol{H}}\mathrm{R}_{ss}\tilde{\boldsymbol{H}}^{\mathcal{H}}$, and its settings are given by: $\mathrm{W}_{mmse} = \mathrm{U}_{mmse}\mathrm{R}_{sy}\mathrm{R}_{yy}^{-1}$, $\mathrm{B}_{mmse} = \mathrm{U}_{mmse} - \mathrm{I}_{4J}$, where U_{mmse} is an upper triangular matrix with unit diagonal given by the Cholesky factorization of $\mathrm{R}_{ss}^{-1} + \tilde{\boldsymbol{H}}^{\mathcal{H}}\mathrm{R}_{nn}^{-1}\tilde{\boldsymbol{H}} = \mathrm{U}_{mmse}^{\mathcal{H}}\mathrm{D}_{mmse}\mathrm{U}_{mmse}$, and D_{mmse} a diagonal matrix. For $N_r = N_t = 1$ (no ST coding) the DF receivers outperform their linear counterparts both in the single user case [9] and in the multiuser case [10].

3.2 Linear Two-stage ST Receiver

The receivers described in Section 3.1 do not exploit the special structure of $\boldsymbol{H}$ in (11). We next propose a linear two-stage ST receiver which takes advantage of the ST code structure. To see how this receiver works, let us define the linear filter $\boldsymbol{W}$ as:

$$\boldsymbol{W}{:=}\begin{pmatrix} \mathrm{I}_{2J} & -\boldsymbol{H}_{1,2}\boldsymbol{H}_{2,2}^{-1} \\ -\boldsymbol{H}_{2,1}\boldsymbol{H}_{1,1}^{-1} & \mathrm{I}_{2J} \end{pmatrix}. \tag{14}$$

Our receiver decodes the information symbols in two steps. First, the linear filter $\boldsymbol{W}$ is applied to $\boldsymbol{y}$ to produce the MUI-free $(\boldsymbol{y}_m^T \ \boldsymbol{y}_\mu^T)^T$ as [c.f. (12)]

$$\begin{pmatrix} \boldsymbol{y}_m \\ \boldsymbol{y}_\mu \end{pmatrix} := \boldsymbol{W}\boldsymbol{y} \overset{(12)}{=} \begin{pmatrix} \boldsymbol{H}_{1,1} - \boldsymbol{H}_{1,2}\boldsymbol{H}_{2,2}^{-1}\boldsymbol{H}_{2,1} & 0_{2J} \\ 0_{2J} & \boldsymbol{H}_{2,2} - \boldsymbol{H}_{2,1}\boldsymbol{H}_{1,1}^{-1}\boldsymbol{H}_{1,2} \end{pmatrix} \tilde{\boldsymbol{s}} + \begin{pmatrix} \boldsymbol{n}_1 \\ \boldsymbol{n}_2 \end{pmatrix}, \tag{15}$$

where $(\boldsymbol{n}_1^T \ \boldsymbol{n}_2^T)^T := \boldsymbol{W}\boldsymbol{n}$. Second, the symbols of users m and μ are recovered from the $2J \times 1$ vectors $\boldsymbol{y}_m$ and $\boldsymbol{y}_\mu$:

$$\boldsymbol{y}_m = \left(\boldsymbol{H}_{1,1} - \boldsymbol{H}_{1,2}\boldsymbol{H}_{2,2}^{-1}\boldsymbol{H}_{2,1}\right) \left(\tilde{\mathrm{s}}_m^T(2n) \ \ \tilde{\mathrm{s}}_m^T(2n+1) \right)^T + \boldsymbol{n}_1 \tag{16a}$$

$$\boldsymbol{y}_\mu = \left(\boldsymbol{H}_{2,2} - \boldsymbol{H}_{2,1}\boldsymbol{H}_{1,1}^{-1}\boldsymbol{H}_{1,2}\right) \left(\tilde{\mathrm{s}}_\mu^T(2n) \ \ \tilde{\mathrm{s}}_\mu^T(2n+1) \right)^T + \boldsymbol{n}_2 \, . \tag{16b}$$

Then, it can be proved that the matrices $\boldsymbol{H}_{1,1} - \boldsymbol{H}_{1,2}\boldsymbol{H}_{2,2}^{-1}\boldsymbol{H}_{2,1}$ and $\boldsymbol{H}_{2,2} - \boldsymbol{H}_{2,1}\boldsymbol{H}_{1,1}^{-1}\boldsymbol{H}_{1,2}$ can be written as:

$$\boldsymbol{H}_{1,1} - \boldsymbol{H}_{1,2}\boldsymbol{H}_{2,2}^{-1}\boldsymbol{H}_{2,1} = \begin{pmatrix} \mathrm{A}_m & \mathrm{B}_m \\ \mathrm{B}_m^* & -\mathrm{A}_m^* \end{pmatrix}, \quad \boldsymbol{H}_{2,2} - \boldsymbol{H}_{2,1}\boldsymbol{H}_{1,1}^{-1}\boldsymbol{H}_{1,2} = \begin{pmatrix} \mathrm{A}_\mu & \mathrm{B}_\mu \\ \mathrm{B}_\mu^* & -\mathrm{A}_\mu^* \end{pmatrix}, \tag{17}$$

where $\mathrm{A}_m, \mathrm{B}_m, \mathrm{A}_\mu, \mathrm{B}_\mu$ are $J \times J$ diagonal matrices. Given (17), we can see that, e.g., $\boldsymbol{y}_m$ of (16a) can be broken into two parts (similar to (5a) and (5b)):

$$\boldsymbol{y}_m = \begin{pmatrix} \mathrm{A}_m & \mathrm{B}_m \\ \mathrm{B}_m^* & -\mathrm{A}_m^* \end{pmatrix} \begin{pmatrix} \tilde{\mathrm{s}}_m(2n) \\ \tilde{\mathrm{s}}_m(2n+1) \end{pmatrix} + \boldsymbol{n}_1 \, . \tag{18}$$

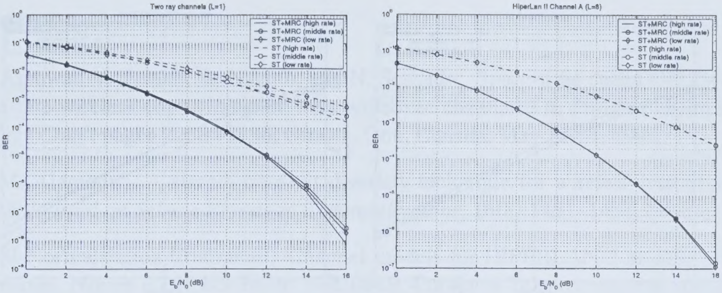
Hence, the symbols of user m can be retrieved with ST diversity gains following the path dictated by (6) and (7).

From a maximum likelihood estimation standpoint, disjoint symbol recovery for users m and μ is not optimal. It can be seen that $\boldsymbol{W}\boldsymbol{W}^{\mathcal{H}}$ is not diagonal, which makes $\boldsymbol{n}$ colored (i.e., $\boldsymbol{n}_1, \boldsymbol{n}_2$ are dependent). Therefore, subcarrier sharing comes with a price to be paid both in the form of suboptimum symbol estimation and in the form of possibly reduced SNR as evinced by the factors $-\boldsymbol{H}_{1,2}\boldsymbol{H}_{2,2}^{-1}\boldsymbol{H}_{2,1}$ and $-\boldsymbol{H}_{2,1}\boldsymbol{H}_{1,1}^{-1}\boldsymbol{H}_{1,2}$ in (16a) and (16b), respectively. We note that the form of $\boldsymbol{W}$ is reminiscent of the matrix linear combiner of [7, eq. (24)], but with a few important differences. First, the pertinent method in [7] assumes flat-fading channels, whereas we take into account the channel order. Second, our block matrices $\boldsymbol{H}_{1,1}, \boldsymbol{H}_{1,2}, \boldsymbol{H}_{2,1}, \boldsymbol{H}_{2,2}$ hold the frequency response of the channels (at the corresponding subcarriers), whereas in [7] the block submatrices hold the fading constants of the flat fading channels.

4 Simulations

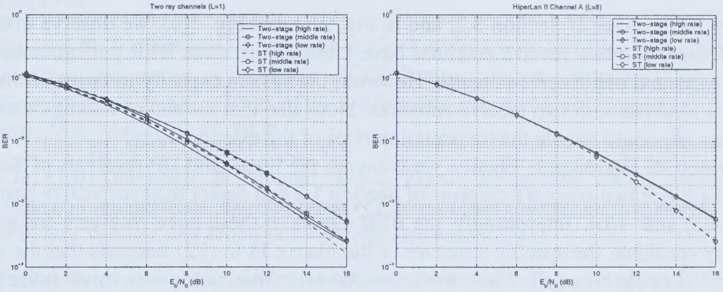
In this section we present simulation results for $N_r = 2$ receive-antennas (see [5, 11] for the $N_r = 1$ case). We study a system with $M = 8$ users, where each user is equipped with two transmit-antennas; two high-rate users have been assigned the weight $\phi_{hr} \approx 1/4$, two middle-rate users have $\phi_{mr} \approx 1/8$, whereas four low-rate users have been assigned the weight $\phi_{lr} \approx 1/16$. In all simulations, the BER is chosen as the figure of merit and is averaged over 600 random channel and noise realizations for various E_b/N_0 points. The random channels are generated based on two different channel models:

Two-ray channel ($L = 1$): This corresponds to the case where the multipath consists of two equal-power dominant rays, and one ray is delayed with respect to the other by a chip duration.



(a) Two-ray channels (L=1) (b) HiperLan 2 channel A (L=8)

Figure 2: Receive diversity gains with MRC($N_r = 2$)



(a) Two-ray channels (L=1) (b) HiperLan 2 channel A (L=8)

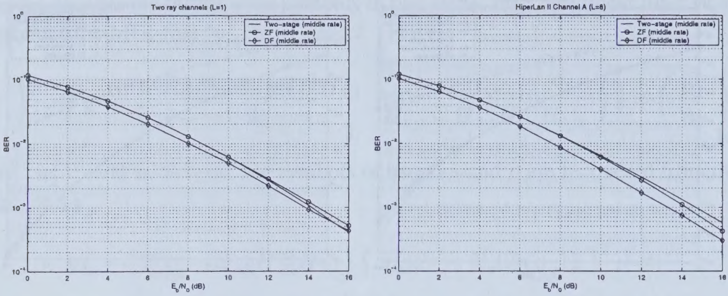
Figure 3: Doubling network capacity with two receive-antennas

HiperLan 2 channel ($L = 8$): The random channels are based on the HiperLan 2 channel model A, which corresponds to a typical office environment [3].

With two receive antennas, there are two alternative paths we can follow to improve the overall system performance. The first (rather well-known) path is to retain the code assignments of Section 2 as if there was only one receive-antenna, and then employ maximum ratio combining (MRC) at each subcarrier to enhance the reliability of the symbol estimates. The second way is to apply subcarrier sharing techniques and then use two antennas to *double* system capacity in terms of either achievable transmission rates or maximum number of users, as we described in Section 3.

Maximum Ratio Combining: Figs. 2(a) and 2(b) illustrate the performance improvement through MRC. It is observed that a 6 dB gain is obtained at BER=10^{-3} as compared to a single receive-antenna. Though it is well known that multiple receive-antennas improve performance, this example indicates that array processing algorithms can be readily applied to our framework.

Doubling Network Capacity: We look at a system with a total of $2M$ users, who are divided into two identical groups with the same subcarrier allocations. In other words, each user in one group shares the same subcarriers with one in the other group. As described in Section 3, two receive-antennas can be used to separate two subcarrier-sharing users. We choose the linear



(a) Two-ray channels (L=1)

(b) HiperLan 2 channel A (L=8)

Figure 4: Comparison of receivers (middle-rate user only)

two-stage ST receiver of Section 3.2, and in Fig. 3(a) and Fig. 3(b), we observe that systems with $2M$ users and two receive-antennas have (almost) the same BER performance as those with M users and one receive-antenna. However, the system capacity in terms of the number of users is doubled with two receive antennas. Note also that instead of adding more users to the system, we could double the transmission rates of the existing users.

Comparison of Receivers: Finally, we compare the linear and DF receivers of Section 3.1 against the linear two-stage ST receiver of Section 3.2 and depict their respective performance in Figs. 4(a) and 4(b). Interestingly, the BER curves of these three receivers are very close. Intuitively speaking, the specific structure of the matrix $\mathbf{H}$ in (12) suggests that the condition number of $\mathbf{H}$ could be small. Thus, the linear receivers may not induce much noise enhancement. Although their BER performance is close, the two-stage ST receiver is better than the other two in terms of computational complexity.

5 Conclusions

In this paper we have developed a multirate/multiuser CDMA system that relies on ST coded block-spread multicarrier transmissions to guarantee deterministic symbol recovery with multipath diversity gains regardless of the (possibly unknown) frequency-selective FIR channels and multiuser interference. Our approach exploits the transmit and possibly receive antenna diversity to improve BER performance and increase (in terms of the number of users or supported transmission rates) the capacity of broadband wireless networks. Furthermore, the physical layer of our system facilitates multirate bandwidth allocation, which is an important feature in Integrated Services networks. From a technical standpoint, our approach is based on a three-level user code design: the outer code handles MUI, the middle code results in ST diversity gains, and the inner code mitigates ISI. Yet, the proposed framework is not computationally expensive, as it relies primarily on FFT-based processing. As illustrated by simulations, the deterministic MUI elimination, ST diversity gains and guaranteed symbol recovery lead to improved BER performance. It is in our plans to study whether it is possible to optimize our framework either in the deterministic sense (when channel status information is available at the transmitter) or in the stochastic sense (when only statistical description of the fading channels is known).

References

[1] D. Agrawal, V. Tarokh, A. Naguib, and N. Seshadri. Space-time coded OFDM for high data-rate wireless communication over wideband channels. In *Proc. of Vehicular Technology Conf.*, pages 2232 –2236, Ottawa, Ont., Canada, May 18-21, 1998.

[2] S. M. Alamouti. A simple transmit diversity technique for wireless communications. *IEEE Journal on Selected Areas in Communications*, 16(8):1451–1458, October 1998.

[3] ETSI Normalization Commitee. Channel models for HIPERLAN/2 in different indoor scenarios. Norme ETSI, document 3ERI085B, European Telecommunications Standards Institute, Sophia-Antipolis, Valbonne, France, 1998.

[4] G. J. Foschini and M. J. Gans. On limits of wireless communication in a fading environment when using multiple antennas. *Wireless Personal Communications*, 6(3):311–335, March 1998.

[5] Z. Liu and G. B. Giannakis. Space-time block coding with transmit antennas for multiple access irrespective of multipath. *IEEE Transactions on Communications*, October 1999 (submitted). See also *Proc. of the 1st Sensor Array and Multichannel SP Workshop*, Boston, MA, March 15-17, 2000.

[6] Z. Liu, G.B. Giannakis, B. Muquet, and S. Zhou. Space-time coding for broadband wireless communications. *Special Issue of International Journal for Computer Research on Wireless Systems and Mobile Computing*, September 2000.

[7] A. F. Naguib, N. Seshadri, and R. Calderbank. Space-time coding and signal processing for high data rate wireless communications. *IEEE Signal Processing Magazine*, 17(3):76–92, May 2000.

[8] A. Scaglione, G. B. Giannakis, and S. Barbarossa. Redundant filterbank precoders and equalizers Part I: Unification and optimal designs. *IEEE Transactions on Signal Processing*, 47:1988–2006, July 1999.

[9] A. Stamoulis, G. B. Giannakis, and A. Scaglione. Block FIR decision-feedback equalizers for filterbank precoded transmissions with blind channel estimation capabilities. *IEEE Transactions on Communications*, 2000 (to appear) see also "Self-recovering Transceivers for Block Transmissions: Filterbank Precoders and Decision-Feedback Equalizers" in *Proc. SPAWC'99*.

[10] A. Stamoulis and G.B. Giannakis. Multiuser decision-feedback equalization of block-spread multirate transmissions for QoS wireless networking. In *Proc. of Wireless Communications and Networking Conf.*, Chicago, Illinois, September 23-28 2000 (to appear); `http://spincom.ece.umn.edu/preprints/wcnc00dfe.ps`.

[11] A. Stamoulis, Z. Liu, and G.B. Giannakis. Supporting Integrated Services in Wireless Networks with Space-Time Block-Coded Transmissions. In *Proc. 34th Annual Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, California, November 2000.

[12] Z. Wang and G. B. Giannakis. Wireless multicarrier communications: Where Fourier meets Shannon. *IEEE Signal Processing Magazine*, 17(3):29–48, May 2000.

Union Bound for Linear Space-Time Codes

Sumeet Sandhu
Packard 225
Stanford University
Stanford, CA 94305-9510
sumeet@stanford.edu

Robert W. Heath Jr.
Packard 234
Stanford University
Stanford, CA 94305-9510
rheath@stanford.edu

A. Paulraj
Packard 272
Stanford University
Stanford, CA 94305-9510
paulraj@rascals.stanford.edu

October 6, 2000

Abstract

Design of practical coding and modulation techniques for the multiple antenna fading wireless channel is a challenging problem. A number of interesting solutions have been proposed recently ranging from block codes to trellis codes for the MIMO (multiple input, multiple output) fading channel. We address the general problem of linear code design for the quasi-static, flat-fading, coherent MIMO channel. A *linear code* refers to an encoder that is linear with respect to scalar input symbols. The decoder performs ML (maximum likelihood) decoding of the received matrix symbols and is not assumed to be linear.

We provide a cohesive framework for analysis of linear codes in terms of a union bound on the *conditional* probability of symbol error. The error bound is conditioned on the channel realization and does not make any assumptions on channel statistics. Our analysis incorporates all existing linear spatial modulation techniques such as spatial multiplexing, space-time block codes, and truncated delay diversity. We show that space-time block codes achieve the lowest error bound among all orthogonal codes and are in fact optimal.

1 Introduction

A number of interesting coding and modulation techniques addressing the problem of transmit diversity design for the multiple-antenna fading channel have been proposed recently. They range from spatial multiplexing [1] to space-time block codes [2] to space-time trellis codes [3]. Among these schemes, the first two are linear and the third is nonlinear, where linearity is with respect to the scalar input symbols that are to be mapped to space-time matrix codewords.

In this paper we address the general problem of linear code design. We define a *linear code* as one that has a linear encoder, i.e., the encoder uses a set of modulation matrices to modulate the sequence of scalar input symbols and adds up the products to create the matrix output codeword. The output codeword is a matrix with dimensions representing transmit antennas and time. The decoder has perfect channel knowledge and performs ML decoding of the matrix symbols, and is not assumed to be linear or otherwise constrained.

We analyze all linear codes within the framework of probability of (matrix) symbol error. We obtain an upper bound on the error probability using the union bound on probabilities. The upper bound is *conditioned on the instantaneous channel realization.* It does not depend on channel statistics and is expected to be tight at high SNRs (signal to noise ratios). Focusing on the class of *orthogonal linear codes*, i.e., where the modulation matrices are orthogonal, we obtain *necessary and sufficient conditions for minimization* of the error bound. In the absence of any channel knowledge at the transmitter these conditions are met by the STBC (space-time block coding) solution. Our analysis does not presuppose decoupled detection as done in [4], leading to it instead in a natural fashion. Our error bound can be used to analyze any linear code, and work is in progress to obtain minimization conditions for all linear codes.

2 Data Model and Notation

Consider a system with M_r receive antennas and $M_t > 1$ transmit antennas. The channel is flat-fading and quasi-static. It is unknown at the transmitter but is known at the receiver. At time nL, the channel output corresponding to the n^{th} input block spanning L symbol times is

$$\mathbf{Y}_{nL} = \mathbf{H}\mathbf{X}_{nL} + \mathbf{V}_{nL} \tag{1}$$

where the received signal $\mathbf{Y}_{nL}$ is $M_r \times L$, the fading channel $\mathbf{H}$ is $M_r \times M_t$, the encoded codeword $\mathbf{X}_{nL}$ is $M_t \times L$, and receiver noise $\mathbf{V}_{nL}$ is $M_r \times L$. The entries of $\mathbf{V}_{nL}$ are i.i.d.(independent, identically distributed), $v_{nL_{i,j}} \sim \aleph_c(0, N_0)$, and independent over n. The average power transmitted on M_t antennas is E_s per symbol time. Define $S = \frac{E_s}{N_0}$.

In our notation $\mathbf{A}^T$ denotes the transpose, $\mathbf{A}^*$ denotes the Hermitian, and Tr$\mathbf{A}$ denotes the trace of matrix $\mathbf{A}$ respectively, and Re denotes the real part. The complementary error function integral is defined as $\mathrm{Q}(x) = \int_x^\infty \frac{dt}{\sqrt{2\pi}} \exp(-\frac{t^2}{2})$.

3 Encoder and Decoder

Following the linear modulation structure in [4], we consider codewords that consist of a set of modulation matrices modulated by scalar input symbols. The input to the encoder is a stream of symbols from a constellation such as PAM, QAM or PSK. Each complex symbol is expanded into two real symbols corresponding to its real and imaginary parts respectively. The encoder operates on the sequence of real symbols producing a matrix codeword whose rows correspond to antennas and columns correspond to symbol times. We define linear codes as follows.

Definition 1 (Linear codes) *A linear code is defined as a set of codewords that are linear in the scalar input symbols. Let $x^{(i)}$ and $x^{(j)}$ be two input symbol sequences with output codewords equal to $\mathbf{X}^{(i)}$ and $\mathbf{X}^{(j)}$ respectively. Then the codeword corresponding to the sequence $\alpha x^{(i)} + \beta x^{(j)}$ is equal to $\alpha\mathbf{X}^{(i)} + \beta\mathbf{X}^{(j)}$, where α, β are any complex scalars. The decoder is the optimal decoder, i.e., it performs ML detection on the observations $\mathbf{Y}$.*

At time nL, the m^{th} block of K input symbols $\{x_{mK-k+1}\}_{k=1}^K$ is used to modulate a set of matrices $\{\mathbf{A}_k\}_{k=1}^K$ which may be real or complex. The block indices n and m may be different depending on the size $M_t \times L$ of matrices used to encode K symbols. For e.g., $K \leq L$ for space-time block codes and $K > L$ for spatial multiplexing. The modulated matrices are summed to obtain the $M_t \times L$ codeword $\mathbf{X}_{nL} = \sum_{k=1}^K \mathbf{A}_k x_{mK-k+1}$, each row of which is then low-pass filtered, D/A sampled, upconverted and transmitted on the corresponding antenna. To simplify notation, we shall drop the time indices and consider the codebook $\mathcal{X}=\{\mathbf{X}^{(1)},\ldots,\mathbf{X}^{(R)}\}$ where the r^{th} codeword $\mathbf{X}^{(r)}$ is constructed from the r^{th} sequence of input symbols $x^{(r)}=\{x_1^{(r)},\ldots,x_K^{(r)}\}$ as follows

$$\mathbf{X}^{(r)}=\sum_{k=1}^{K}\mathbf{A}_k x_k^{(r)} \tag{2}$$

Depending on the input symbol constellation, each real symbol $x_k^{(r)}$ belongs to a PAM, one-dimensional "QAM" or one-dimensional "PSK" constellation of size M. The average energy of the PAM constellation is E_s but the average energy of the real and imaginary symbols from the QAM and PSK constellations is $\frac{E_s}{2}$ each. The total number of codewords is $R=|\mathcal{X}|=\mathrm{M}^K$. We assume no additional outer coding on the input symbols in the subsequent analysis and comment on the effect of coded symbols in the conclusion. Since we consider the real and imaginary parts of a complex input symbol as two separate input symbols, the codeword in (2) is *linear in the input symbols* over the complex field.

We impose a power constraint on each modulation matrix by requiring $\|\mathbf{A}_k\|_F^2 \le c$ for all k. This ensures that each codeword is bounded in power as follows

$$\begin{aligned}
\|\mathbf{X}\|_F^2 &= \|\sum_{k=1}^{K}\mathbf{A}_k x_k\|_F^2 = \sum_{i=1}^{M_t}\sum_{j=1}^{L}|\sum_{k=1}^{K}\mathbf{A}_{k[i,j]}x_k|^2 \\
&\le \sum_{i=1}^{M_t}\sum_{j=1}^{L}\sum_{k=1}^{K}|\mathbf{A}_{k[i,j]}x_k|^2 = \sum_{k=1}^{K}|x_k|^2\sum_{i=1}^{M_t}\sum_{j=1}^{L}|\mathbf{A}_{k[i,j]}|^2 \qquad (3)\\
&= \sum_{k=1}^{K}|x_k|^2\|\mathbf{A}_k\|_F^2 \le c\sum_{k=1}^{K}|x_k|^2 \qquad (4)
\end{aligned}$$

where (3) follows from the Cauchy-Schwartz inequality, and (4) follows from the power constraint $\|\mathbf{A}_k\|_F^2 \le c$. From (4) we observe that the average power of the matrix codewords is at most $\frac{cK}{L}E_s$ per symbol time for PAM input constellations because

$$\frac{1}{L}\mathrm{E}\|\mathbf{X}\|_F^2 \;\le\; \frac{c}{L}\mathrm{E}\sum_{k=1}^{K}|x_k|^2 \;=\; \frac{cK}{L}\mathrm{E}|x_k|^2 \;=\; \frac{cK}{L}E_s$$

In order to maintain the average power at E_s per symbol time, we must normalize the modulation matrices for different input constellations as follows

$$\begin{aligned}
&\text{PAM} && \|\mathbf{A}_k\|_F^2 \le \frac{L}{K} \text{ for all } k \\
&\text{QAM , PSK} && \|\mathbf{A}_k\|_F^2 \le \frac{2L}{K} \text{ for all } k
\end{aligned} \tag{5}$$

We assume a coherent receiver, i.e., the channel is perfectly known. The received $M_r{\times}L$ signal $\mathbf{Y}$ is decoded using maximum likelihood decoding. Some practical examples of linear codes are presented in the next section to clarify concepts.

4 Examples of linear codes

In order to clarify the concept of linear codes, we provide a few examples in this section. Most of the current spatial modulation techniques can be considered as linear codes, the exceptions being space-time trellis codes.

Example 1 (Spatial multiplexing) *Spatial multiplexing [1], also called BLAST [5], is the simplest example of linear codes. Each incoming symbol is transmitted once on one antenna only. The modulation matrices are simply unit vectors, i.e., the k^{th} matrix is an $M_t \times 1$ vector that is nonzero in the k^{th} position. For e.g., for a PAM input constellation, $A_k = \frac{1}{\sqrt{M_t}}[0 \cdots 1 \cdots 0]^T$. For M_t transmit antennas, $K = M_t$ and $L = 1$.*

Example 2 (Space-time block coding) *Space-time block coding is another example of linear codes. The modulation matrices are unitary matrices that are pairwise orthogonal with respect to the* Re Tr *matrix inner product. That is, for any weighting matrix* $\mathbf{W}$, $\mathrm{Re}\,\mathrm{Tr}(\mathbf{A}_k^*\mathbf{W}\mathbf{A}_l) = 0$ *for all* $1 \leq k \neq l \leq K$*. Such matrices exist for limited values of* K*,* M_t *and* L *[2]. For e.g., the complex, full rate Alamouti code [6] has* $M_t = 2$ *transmit antennas,* $K = 4$ *real symbols (from 2 complex input symbols), and* $L = 2$*. The two real, full rate, minimum-delay Tarokh codes have* $M_t = K = L = 4$ *and 8 respectively.*

Example 3 (Delay diversity) *Delay diversity [7, 8] can be considered as a linear code if the tail effects are ignored ("truncated" delay diversity). The* $M_t \times L$ *modulation matrices are proportional to* $[0 I_{M_t} 0]$*, where the order* M_t *identity matrix is shifted right by* $k-1$ *columns in the* k^{th} *modulation matrix. For any* $L \geq M_t$ *the total number of scalar symbols thus encoded is* $K = L - M_t + 1$*. In order that the loss in rate by ignoring the tail symbols be negligible,* L *must be large.*

Example 4 (Cyclic delay diversity) *We propose a* new *class of linear codes called cyclic delay diversity. Instead of being dropped altogether, the tail symbols transmitted at the beginning of the block are repeated at the end of the block. This results in cyclic modulation matrices and an equivalent channel that is circulant. Similar to OFDM (Orthogonal Frequency Division Multiplexing) over Toeplitz channel matrices, encoding/decoding becomes much simpler because IFFT (Inverse Fast Fourier Transform) transmit matrices and FFT receive matrices diagonalize the equivalent circulant channel. Performance analysis of this scheme is currently in progress.*

5 Performance Analysis

We will analyze linear codes using the conditional probability of (matrix) symbol error, i.e., the error probability based on a given realization of the channel which is known at the receiver. First we compute an upper bound on the probability of symbol error in Lemma 1, then we set up conditions for its minimization in Lemmas 2 and 3. We state necessary and sufficient conditions for minimization over orthogonal codes in Lemma 4. Finally in Lemma 5 we show that the optimal linear code is the space-time block code, which is stated as the main result of the paper in Theorem 1.

Lemma 1 *The probability of matrix symbol error can be bounded in the following manner*

$$P_e \;\leq\; \sum_{i=1}^{R} p_i \sum_{j \neq i}^{R} \mathrm{Q}\left(\sqrt{\Delta_1^{(ij)} + \Delta_2^{(ij)}}\right) \;=\; P_U \tag{6}$$

where the argument of each Q-function is a function of two terms, one of which $\Delta_1^{(ij)}$ *is determined by the norms of individual modulation matrices and the other* $\Delta_2^{(ij)}$ *is determined by their pairwise inner products.*

Proof: Lemma 1 is proved by construction. The probability of matrix symbol error is the average of the pairwise error probabilities over all matrix symbols as follows

$$P_e \;=\; \sum_{i=1}^{R} p_i P_{e|i} \tag{7}$$

where p_i is the probability that codeword i is transmitted, and $P_{e|i}$ is the probability that the receiver does not decode i correctly. The probability of detecting i incorrectly is equal to the probability that one of the other codewords j is detected where $j \neq i$. Using the union bound on probabilities, $P_{e|i}$ can be bounded as follows

$$P_{e|i} \;\leq\; \sum_{j \neq i}^{R} \mathrm{Q}\left(\sqrt{D_{ij} S/2}\right). \tag{8}$$

For a given channel realization $\mathbf{H}$, D_{ij} is the pairwise Euclidean distance at the receiver defined as

$$D_{ij} \;=\; \|\mathbf{H}\,(\mathbf{X}^{(i)} - \mathbf{X}^{(j)})\|_F^2 \;=\; \|\mathbf{H} \sum_{k=1}^{K} \mathbf{A}_k (x_k^{(i)} - x_k^{(j)})\|_F^2 = \|\mathbf{H} \sum_{k=1}^{K} \mathbf{A}_k \delta_k^{(ij)}\|_F^2 \tag{9}$$

where $\delta_k^{(ij)} = x_k^{(i)} - x_k^{(j)}$ is the difference between sequences i and j at the k^{th} position. It can be further rewritten as

$$\begin{aligned}
D_{ij} \;&=\; \mathrm{Tr}\left(\left[\mathbf{H} \sum_{k=1}^{K} \mathbf{A}_k \delta_k^{(ij)}\right]\left[\sum_{l=1}^{K} \mathbf{A}_l^* \mathbf{H}^* (\delta_l^{(ij)})^*\right]\right) &(10)\\
&=\; \mathrm{Tr}\left(\sum_{k=1}^{K}\sum_{l=1}^{K} \mathbf{H}\mathbf{A}_k\mathbf{A}_l^*\mathbf{H}^* \delta_k^{(ij)} (\delta_l^{(ij)})^*\right) &(11)\\
&=\; \mathrm{Tr}\left(\sum_{k=1}^{K} \mathbf{H}\mathbf{A}_k\mathbf{A}_k^*\mathbf{H}^* \delta_k^{(ij)} (\delta_k^{(ij)})^*\right) + \mathrm{Tr}\left(\sum_{k \neq l}^{K} \mathbf{H}\mathbf{A}_k\mathbf{A}_l^*\mathbf{H}^* \delta_k^{(ij)} (\delta_l^{(ij)})^*\right) &(12)\\
&=\; \sum_{k=1}^{K} \|\mathbf{H}\mathbf{A}_k\|_F^2 |\delta_k^{(ij)}|^2 + 2\sum_{k<l}^{K} \mathrm{Re}\;\mathrm{Tr}\left(\mathbf{H}\mathbf{A}_k\mathbf{A}_l^*\mathbf{H}^* \delta_k^{(ij)} (\delta_l^{(ij)})^*\right) &(13)
\end{aligned}$$

Here (10) follows from (9) because $\|\mathbf{H}\|_F^2 = \mathrm{Tr}(\mathbf{H}\mathbf{H}^*)$ [9], (11) and (12) are simple rearrangments of the terms in (10), and (13) uses the fact that $a + a^* = 2\mathrm{Re}(a)$. Invoking the fact that the input symbols and their differences $\delta_k^{(ij)}$ are real, we can write D_{ij} as

$$D_{ij} \;=\; \sum_{k=1}^{K} \Omega_{kk} |\delta_k^{(ij)}|^2 + 2\sum_{k<l}^{K} \Omega_{kl} \delta_k^{(ij)} \delta_l^{(ij)} \tag{14}$$

Using the property that $\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$ [9], $\Omega_{kl} = \mathrm{Re}\;\mathrm{Tr}(\mathbf{H}\mathbf{A}_k\mathbf{A}_l^*\mathbf{H}^*)$ is the symmetric, weighted, matrix inner product of $\mathbf{A}_k$ and $\mathbf{A}_l$ defined as

$$\Omega_{kl} \;=\; <\mathbf{A}_k, \mathbf{A}_l>_{\mathbf{H}^*\mathbf{H}} \;=\; \frac{1}{2}(\mathrm{Tr}(\mathbf{A}_k^*\mathbf{H}^*\mathbf{H}\mathbf{A}_l) + \mathrm{Tr}(\mathbf{A}_l^*\mathbf{H}^*\mathbf{H}\mathbf{A}_k)) \;=\; \mathrm{Re}\;\mathrm{Tr}(\mathbf{A}_k^*\mathbf{H}^*\mathbf{H}\mathbf{A}_l)$$

where the weighting matrix is $\mathbf{H}^*\mathbf{H}$. Note that $\Omega_{kk} = \|\mathbf{H}\mathbf{A}_k\|_F^2 \geq 0$. Define the following

$$\begin{aligned}
\Delta_1^{(ij)} \;&\triangleq\; \frac{S}{2}\sum_{k=1}^{K} \Omega_{kk} \left|\delta_k^{(ij)}\right|^2 \;=\; \frac{S}{2}\sum_{k=1}^{K} \|\mathbf{H}\mathbf{A}_k\|_F^2 \left|\delta_k^{(ij)}\right|^2, \\
\Delta_2^{(ij)} \;&\triangleq\; \frac{S}{2}\, 2\sum_{k<l}^{K} \Omega_{kl} \delta_k^{(ij)} \delta_l^{(ij)} \;=\; S\sum_{k<l}^{K} \mathrm{Re}\;\mathrm{Tr}(\mathbf{A}_k^*\mathbf{H}^*\mathbf{H}\mathbf{A}_l) \delta_k^{(ij)} \delta_l^{(ij)} &(15)
\end{aligned}$$

Of these, $\Delta_1^{(ij)}$ is a function of the weighted norms of individual modulation matrices, and $\Delta_2^{(ij)}$ is a function of the weighted pairwise inner products of the modulation matrices.

Substituting (15) in (13), (8) and then in (7) we get

$$P_e \;\le\; \sum_{i=1}^{R} p_i \sum_{j\neq i}^{R} \mathrm{Q}\left(\sqrt{D_{ij}S/2}\right) \;=\; \sum_{i=1}^{R} p_i \sum_{j\neq i}^{R} \mathrm{Q}\left(\sqrt{\Delta_1^{(ij)}+\Delta_2^{(ij)}}\right) \;=\; P_U \tag{16}$$

which finishes the proof of Lemma 1. ■

The goal of code design is to find modulation matrices $\{\mathbf{A}_k\}_{k=1}^{K}$ that minimize the upper bound on the conditional probability of symbol error P_U. The complementary error function (Q-function) in (16) is a monotonically decreasing function of its argument. In order to minimize P_U, we must maximize the argument of each of the Q-functions. From (16), the argument of each Q-function $\sqrt{D_{ij}S/2}$ is a function of two terms, one of which is determined by the norms of individual modulation matrices Ω_{kk} and the other is determined by their pairwise inner products Ω_{kl}. If we make the assumption that all sequences are equally likely, i.e., $p_i = \frac{1}{R}$ for all i in (16), then we have Lemma 2.

Lemma 2 *: By carefully selecting terms over i and j, we can always group terms in the expression for the upper bound on the probability of symbol error into pairs of the following form*

$$q \;=\; \frac{n}{R}\left[\mathrm{Q}\left(\sqrt{\Delta_1+\Delta_2}\right)+\mathrm{Q}\left(\sqrt{\Delta_1-\Delta_2}\right)\right] \tag{17}$$

where $\Delta_1 \ge \Delta_2 \ge 0$ *and n is an integer denoting the number of such pairs.*

Proof: Recall that the i^{th} input sequence consists of K symbols, i.e., $x^{(i)} = \{x_1^{(i)},\ldots,x_K^{(i)}\}$. The difference between $x^{(i)}$ and $x^{(j)}$ is the sequence $\delta^{(ij)} = \{\delta_1^{(ij)},\ldots,\delta_K^{(ij)}\}$, where $\delta_k^{(ij)} = x_k^{(i)} - x_k^{(j)}$. When two symbol sequences differ only in m positions, say $k_1,\ldots,k_m$, then $\delta_k^{(ij)} \neq 0$ for $k = k_1,\ldots,k_m$, and $\delta_k^{(ij)} = 0$ for $k \neq k_1,\ldots,k_m$. Let $\delta_{k_i}^{(ij)} = \delta_i \neq 0$ for $1 \le i \le m$. The range of values of δ_i is determined by the input constellation (PAM, QAM or PSK). The upper bound in (16) is a sum of contributions from difference sequences corresponding to all possible values of δ_i, k_i and m.

Example 5 *As an example we illustrate the binary constellation with two modulation matrices in Table 1. The rows represent the* i^{th} *sequence and the columns represent the*

$x^{(i)}$ / $x^{(j)}$	(-1 -1)	(-1 1)	(1 -1)	(1 1)
(-1 -1)	(0 0)	(0 -2)	(-2 0)	(-2 -2)
(-1 1)	(0 2)	(0 0)	(-2 2)	(-2 0)
(1 -1)	(2 0)	(2 -2)	(0 0)	(0 -2)
(1 1)	(2 2)	(2 0)	(0 2)	(0 0)

Table 1: All difference sequences for M = 2, K = 2

j^{th} *sequence. The* $[ij]$ *entry represents* $\delta^{(ij)}$ *which is a length two difference sequence.*

The total number of sequences is $M^K = 4$. *Computing (16) using Table 1, we have*

$$\begin{aligned} P_U &= Q\left(\sqrt{\frac{S}{2}\Omega_{11}(2)^2}\right) + Q\left(\sqrt{\frac{S}{2}\Omega_{22}(2)^2}\right) \ldots \\ &\quad + \frac{1}{2}Q\left(\sqrt{\frac{S}{2}(\Omega_{11}(2)^2 + \Omega_{22}(2)^2) + S\Omega_{12}(2)(2)}\right) \ldots \\ &\quad + \frac{1}{2}Q\left(\sqrt{\frac{S}{2}(\Omega_{11}(2)^2 + \Omega_{22}(2)^2) - S\Omega_{12}(2)(2)}\right) \end{aligned} \tag{18}$$

where the first two terms satisfy Lemma 2 trivially with $\Delta_1 = \frac{S}{2}\Omega_{11}(2)^2$ *and* $\Delta_2 = 0$, *and* $\Delta_1 = \frac{S}{2}\Omega_{22}(2)^2$ *and* $\Delta_2 = 0$ *respectively. The last two terms satisfy Lemma 2 with* $\Delta_1 = \frac{S}{2}(\Omega_{11}(2)^2 + \Omega_{22}(2)^2)$ *and* $\Delta_2 = |S\Omega_{12}(2)(2)|$.

The key concept for the general proof of Lemma 2 follows from the simple case $m=1$, i.e., those pairs of sequences that differ in only one position. Consider all difference sequences such that $|\delta_k^{(ij)}| = \delta_1$ for $k = k_1$ and zero otherwise. There are two possible signs for the actual difference, either $\delta_{k_1}^{(ij)} = \delta_1$ or $\delta_{k_1}^{(ij)} = -\delta_1$. By symmetry, the number of difference sequences with $\delta_{k_1}^{(ij)} = \delta_1$ is the same as the number of difference sequences with $\delta_{k_1}^{(ij)} = -\delta_1$, say n_1 each [1]. Since only one difference position is nonzero in (15), $\Delta_2^{(ij)} = S\sum_{k<l}^{K}\Omega_{kl}\delta_k^{(ij)}\delta_l^{(ij)} = 0$. For both signs of δ_1, $\Delta_1^{(ij)} = \frac{S}{2}\Omega_{k_1k_1}\delta_1^2$. Since *all sequences are equally likely*, the contribution to P_U from these $2n_1$ difference sequences is

$$q = \frac{n_1}{R}\left[Q\left(\sqrt{\Delta_1^{(ij)}+0}\right) + Q\left(\sqrt{\Delta_1^{(ij)}+0}\right)\right] = \frac{2n_1}{R}Q\left(\sqrt{\frac{S}{2}\Omega_{k_1k_1}\delta_1^2}\right)$$

This pair satisfies Lemma 2 trivially with $\Delta_1 = \Delta_1^{(ij)}$ and $\Delta_2 = 0$.

For the general proof, consider the Euclidean distance D_m between two sequences when they differ in m positions $k_1, \ldots, k_m$ by absolute values $|\delta_1|, \ldots, |\delta_m|$ respectively

$$D_m = (\Omega_{k_1k_1}\delta_1^2 + \ldots + \Omega_{k_mk_m}\delta_m^2) + 2(\Omega_{k_1k_2}\delta_1\delta_2 + \ldots + \Omega_{k_{m-1}k_m}\delta_{m-1}\delta_m)$$

Now consider the distance when m is increased to $m+1$.

$$\begin{aligned} D_{m+1} &= (\Omega_{k_1k_1}\delta_1^2 + \ldots + \Omega_{k_{m+1}k_{m+1}}\delta_{m+1}^2) + 2(\Omega_{k_1k_2}\delta_1\delta_2 + \ldots + \Omega_{k_mk_{m+1}}\delta_m\delta_{m+1}) \\ &= D_m + \Omega_{k_{m+1}k_{m+1}}\delta_{m+1}^2 + 2(\Omega_{k_1k_{m+1}}\delta_1 + \ldots + \Omega_{k_mk_{m+1}}\delta_m)\delta_{m+1} \end{aligned} \tag{19}$$

Similar to the case $m=1$, the actual difference at position k_{m+1} is either δ_{m+1} or $-\delta_{m+1}$ with equal probability. For a fixed value of D_m, the argument of the Q function is either $\frac{S}{2}D_{m+1} = \Delta_1 + \Delta_2$ or $\frac{S}{2}D_{m+1} = \Delta_1 - \Delta_2$, where $\Delta_1 = \frac{S}{2}(D_m + \Omega_{k_{m+1}k_{m+1}}\delta_{m+1}^2) \geq 0$ and $\Delta_2 = S|(\Omega_{k_1k_{m+1}}\delta_1 + \ldots + \Omega_{k_mk_{m+1}}\delta_m)\delta_{m+1}| \geq 0$. For a given D_m then, the contribution from $2n_{m+1}$ sequences with differences $\pm\delta_{m+1}$ at position k_{m+1} can be written as

$$q = \frac{n_{m+1}}{R}\left[Q\left(\sqrt{\Delta_1 + \Delta_2}\right) + Q\left(\sqrt{\Delta_1 - \Delta_2}\right)\right]$$

Since this holds for all possible values of δ_i, k_i, $1 \leq i \leq m$, and m, we conclude that all terms in (16) satisfy Lemma 2. ■

---

[1] If there is a sequence pair (i,j) with $x_{k_1}^{(i)} - x_{k_1}^{(j)} = \delta_1$, then by symmetry there is also the pair (j,i) such that $x_{k_1}^{(j)} - x_{k_1}^{(i)} = -\delta_1$.

Lemma 2 sets the stage for minimization of P_U. We will prove Lemma 3 before stating sufficient conditions for minimization in Lemma 4.

Lemma 3 *: For a given value of* Δ_1*, the pair* $q = \mathrm{Q}\left(\sqrt{\Delta_1+\Delta_2}\right) + \mathrm{Q}\left(\sqrt{\Delta_1-\Delta_2}\right)$ *in (17) is minimized iff (if and only if)* $\Delta_2 = 0$.

Proof: This lemma follows from convexity of the Q-function. Since $\mathrm{Q}(\sqrt{x})$ is convex and nonincreasing in $\sqrt{x}$, *and* $\sqrt{x}$ is concave in x for $x > 0$, it follows that $\mathrm{Q}(\sqrt{x})$ is convex in x for $x > 0$ [10]. Applying Jensen's Inequality [11] to q we get Lemma 3. ■

Applying the results of Lemmas 2 and 3 to Lemma 1, we state conditions for minimization of P_U in Lemma 4, subject to structure on the modulation matrices which is described next.

Consider the case $K = 1$, i.e. the linear code consists of only one modulation matrix $\mathbf{A}_1$. This case provides the *MFB (matched filter bound) performance* of the code. The error sequence is at most length one. The error bound P_U consists of terms of the form $\mathrm{Q}(\sqrt{\Delta_1})$ where $\Delta_1 = \frac{S}{2}\Omega_{11}\delta_1^2$ for all possible values of δ_1. In order to minimize P_U it is sufficient to maximize $\Omega_{11} = \|\mathbf{HA}_1\|_F^2$, which can be expanded as follows

$$\begin{aligned} \|\mathbf{HA}_1\|_F^2 &= \mathrm{Tr}(\mathbf{HA}_1\mathbf{A}_1^*\mathbf{H}^*) = \mathrm{Tr}(\mathbf{HU\Sigma\Sigma^*U^*H^*}) = \mathrm{Tr}(\tilde{\mathbf{H}}\mathbf{\Sigma\Sigma}^*\tilde{\mathbf{H}}^*) && (20) \\ &= \mathrm{Tr}\Big(\sum_{i=1}^{\min(M_t,L)} \sigma_i^2\tilde{\mathbf{h}}_i\tilde{\mathbf{h}}_i^*\Big) = \sum_{i=1}^{\min(M_t,L)} \sigma_i^2\mathrm{Tr}(\tilde{\mathbf{h}}_i\tilde{\mathbf{h}}_i^*) = \sum_{i=1}^{\min(M_t,L)} \sigma_i^2\|\tilde{\mathbf{h}}_i\|^2 && (21) \end{aligned}$$

Here (20) follows from the s.v.d.(singular value decomposition) of the $M_t \times L$ matrix $\mathbf{A}_1$ which is $\mathbf{A}_1 = \mathbf{U\Sigma V}^*$, and $\{\sigma_i\}_{i=1}^{\min(M_t,L)}$ are the singular values of $\mathbf{A}_1$. The outer product decomposition of the matrix product is used in (21) where $\tilde{\mathbf{h}}_i$ are the columns of $\mathbf{HU}$. Since the channel is unknown at the transmitter, it is advisable to choose $\sigma_i^2 \neq 0$ for all i in order to maximize the sum in (21). Since the values of $\|\tilde{\mathbf{h}}_i\|^2$ are also unknown, it advisable to choose all the singular values to be equal, i.e., $\sigma_i^2 = \frac{c}{\min(M_t,L)}$ where c is the Frobenius norm of $\mathbf{A}_1$ from (5). This results in $\mathbf{A}_1$ being a unitary matrix.

This result was also obtained in [4] in a similar manner. The optimality of this solution may be also derived using game theory, and we believe that this is the minimax saddlepoint solution. From a stochastic viewpoint, we believe that this solution is optimal when $\mathbf{H}$ is spatially uniform, i.e., it has no preferred directions along its right singular vectors. If $\mathbf{H}$ has an arbitrary mean or covariance, then unitary matrices may not be optimal. Rigorous proof of this is in progress. In the sequel we will assume all matrices are orthogonal, i.e., $\mathbf{A}_k\mathbf{A}_k^* = I_{M_t}$, where $M_t \leq L$, such that

$$\Omega_{kk} = \|\mathbf{HA}_k\|_F^2 = \frac{c}{\min(M_t, L)}\|\mathbf{H}\|_F^2 \qquad \text{for } 1 \leq k \leq K \qquad (22)$$

where c is the power normalization as in (5).

Lemma 4 *A linear code consisting of orthogonal modulation matrices* $\{\mathbf{A}_k\}_{k=1}^K$ *achieves the minimum* P_U *iff the matrices satisfy the following condition*

$$\Omega_{kl} = \frac{1}{2}(\mathrm{Tr}(\mathbf{A}_k^*\mathbf{H}^*\mathbf{HA}_l) + \mathrm{Tr}(\mathbf{A}_l^*\mathbf{H}^*\mathbf{HA}_k)) = 0 \qquad \text{for } 1 \leq k \neq l \leq K \qquad (23)$$

In other words, the modulation matrices must be *pairwise orthogonal* for any weight matrix $\mathbf{H}^*\mathbf{H}$.

Proof: We will prove this lemma by induction on K, the number of modulation matrices. To verify the initial case, consider $K=1$, i.e. the linear code consists of only one modulation matrix $\mathbf{A}_1$. This trivially satisfies (23).

Now we make the inductive assumption, i.e., the optimal set of modulation matrices satisfies Lemma 4 when $K=K'$. Then P_U consists of terms of the following form

$$\begin{aligned} q &= \frac{n_m}{M^K} \mathrm{Q}(\sqrt{\Delta_1}) \\ \text{where} \quad \Delta_1 &= \frac{S}{2} \sum_{i=1}^{m} \Omega_{k_i k_i} \delta_i^2 \end{aligned}$$

where $1 \leq m \leq K'$ with $\Omega_{k_i k_i}$ as in (22).

When the $(K+1)^{th}$ modulation matrix is introduced, we know from (19) that it adds terms of the following form to P_U

$$\begin{aligned} q &= \frac{n}{M^{K+1}} \left[\mathrm{Q}\left(\sqrt{\Delta_1 + \Delta_2}\right) + \mathrm{Q}\left(\sqrt{\Delta_1 - \Delta_2}\right) \right] \qquad (24) \\ \text{where} \quad \Delta_1 &= \frac{S}{2} \sum_{i=1}^{m} \Omega_{k_i k_i} \delta_i^2 + \frac{S}{2} \Omega_{K'+1 K'+1} \delta_{K'+1}^2 \,, \\ \Delta_2 &= S \Big| \sum_{i=1}^{m} \Omega_{k_i K'+1} \delta_i \Big| \left| \delta_{K'+1} \right| \end{aligned}$$

and there are n sequences with a given absolute value of $\delta_{K'+1}$ when all other differences $\{\delta_1, \ldots, \delta_m\}$ are held constant. From Lemma 3 we know that for a fixed value of Δ_1, (24) will be minimized iff $\Delta_2=0$. Since all modulation matrices are assumed to be orthogonal, $\Omega_{K'+1K'+1}$ is as in (22) and Δ_1 is not a function of the actual $(K+1)^{th}$ matrix as long as $\mathbf{A}_{K'+1}$ is orthogonal.

The value of Δ_2 is a function of $\Omega_{k_i K'+1}$, $1 \leq i \leq m$, the weighted inner products of matrices $\mathbf{A}_{k_i}$ with the new matrix $\mathbf{A}_{K'+1}$. The necessary and sufficient condition to ensure $\Delta_2=0$ for all values of k_i, $1 \leq i \leq m$ and m is that $\Omega_{k_i K'+1} = 0$ for $1 \leq i \leq m$. This can be seen from the following system of linear equations that the matrix of all possible differences δ and the vector of pairwise inner products Ω must satisfy

$$\delta \Omega = \mathbf{0}$$

The matrix δ is $m^q \times m$, where m^q is the number of all possible values that m differences can take. The vector Ω is $m \times 1$ and consists of the inner products $\Omega_{k_i K'+1}$ for $1 \leq i \leq m$. This equation must be satisfied for all values of $1 \leq m \leq K'$. Note that δ is a full rank matrix for any m^q, therefore $\Omega=\mathbf{0}$ for all m. This proves that $K'+1$ matrices must also satisfy (23), thereby finishing the induction and the proof of Lemma 4. ■

Lemma 4 takes us to the main result of this paper, which is Theorem 1.

Theorem 1 *Among all orthogonal, linear space-time codes, space-time block codes minimize the union bound on the conditional probability of symbol error for equally likely input symbols.*

Proof: Space-time block codes [2] consist of orthogonal matrices that satisfy Lemma 4. Therefore they satisfy the sufficient and necessary condition for minimization of P_U. In the complete absence of channel knowledge at the transmitter, these are also the only orthogonal codes that minimize the union bound. ■

6 Conclusions

We have provided a new criterion for performance evaluation of linear space-time codes, namely the union bound on symbol error probability. This bound is a function of the instantaneous channel realization and is independent of channel statistics. It is expected to be tight at high SNRs and can be used to analyze average performance over a stochastic channel.

We provided necessary and sufficient conditions for minimization of the union bound over all orthogonal, linear space-time codes. We showed that space-time block codes satisfy these conditions and are therefore optimal. Our analysis applies to equally likely input symbols and does not necessarily imply that space-time block codes are optimal for non equally likely input symbols. We considered modulation matrices for the case when the number of transmit antennas is smaller than the code block length. Extension to modulation matrices when block length is smaller than the number of transmit antennas is in progress, as is extension of our analysis to all linear codes.

References

[1] A. Paulraj and T. Kailath, "Increasing capacity in wireless broadcast systems using distributed transmission/directional reception (DTDR)," Tech. Rep. U.S. Patent #5345599, September 1994.

[2] V. Tarokh, H. Jafarkhani, and A. R. Calderbank, "Space-time block codes from orthogonal designs," *IEEE Trans. Information Theory*, vol. 45, no. 5, pp. 1456–1467, 1999.

[3] V. Tarokh, N. Seshadri, and A. R. Calderbank, "Space-time codes for high data rate wireless communication I: performance criterion and code construction," *IEEE Trans. Information Theory*, vol. 44, pp. 744–65, March 1998.

[4] G. Ganesan and P. Stoica, "Space-time diversity using orthogonal and amicable orthogonal designs," in *Proc. ICASSP*, 2000.

[5] G. J.Foschini and M. J. Gans, "On limits of wireless communications in a fading environment when using multiple antennas," *Wireless Personal Comm.*, vol. 6, pp. 311–335, March 1998.

[6] S. M. Alamouti, "A simple transmitter diversity scheme for wireless communications," *IEEE J. Select. Areas in Comm.*, vol. 16, pp. 1451–1458, October 1998.

[7] A. Wittneben, "Basestation modulation diversity for digital simulcast," in *Vehicular Technology Conference*, pp. 848–853, 1991.

[8] N. Seshadri and J. H. Winters, "Two signaling schemes for improving the error performance of frequency-division-duplex (FDD) transmission systems using transmitter antenna diversity," in *Vehicular Technology Conference*, pp. 508–11, 1993.

[9] H. Lutkepohl, *Handbook of Matrices.* Wiley, 1996.

[10] S. Boyd and L. Vandenberghe, "Convex Optimization." Course notes, 1996-97.

[11] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* Wiley, 1991.

A Random Matrix Model of Communication Through Antenna Arrays

Ralf R. Müller
Vienna Research Center for Telecommunications (FTW)
Maderstr. 1/9, 1040 Vienna, Austria
mueller@ftw.at

Abstract

A random matrix model is introduced that probabilistically describes the multi–path propagation between a transmitting and receiving antenna array with a limited number of scatterers for mobile radio and indoor environments. The model characterizes the channel by its *scatterer count delay profile*. Each delay is assigned the eigenvalue distribution of a random matrix that depends on the number of receive antennas, transmit antennas, and scatterers.

The model allows to calculate signal–to–interference–and–noise ratios and channel capacities for large antenna arrays analytically and shows to which extend rich scattering improves performance.

1 Introduction

Communication through antenna arrays allows for a significant increase in spectral efficiency [1, 2]. While several recent proposals [1, 3, 4, 5] are aiming to utilize this advantage, it is still hardly understood how physical properties of those channels translate into achievable *signal–to–interference–and–noise ratio*s (SINRs) and capable information rates: Physical channel models are based on propagation measurements. They provide statistics of the propagation between a pair of transmitter and receiver arrays in terms of delays, received powers, and directions of arrival [6]. In contrast, the capable information rates of the channel are given in terms of the eigenvalues related to the matrix–algebraic description of the communication link. This work aims to provide the missing link between propagation statistics and the eigenvalues of the channel matrices, and to allow for predictions of channel capacity based on the morphology of the physical medium.

Random matrix theory was used for analysis of antenna arrays in [7, 8, 9]. These references have not taken into account the inter–symbol interference caused by various time delays. In addition, they assumed the entries of the channel matrix to be statistically independent. This is not justified when the number of scattering objects does not significantly exceed both the number of transmit and receive antennas.

Section 2 introduces a channel model in matrix–algebraic notation that accounts for dependencies among the matrix entries caused by limited supply of scattering objects as well as inter–symbol interference. Section 3 shows how to cope with these dependencies in random matrix theory and finds statistics of the eigenvalues of the channel. They are used to derive asymptotic expressions for channel capacity and the SINR that is achievable with ideal linear beam forming at receiver site. Finally, Section 4 summarizes the main results and points out the conclusions.

2 Channel Model

Consider a communication link with T transmit and R receive antennas. Let there be $S_{\max}$ scattering objects each corresponding to a propagation path with excess delay τ_κ. Then, the signal that is received at antenna ν is given by

$$y_\nu(t) = \sum_{\kappa=1}^{S_{\max}} A_\kappa \mathrm{e}^{\mathrm{j}\varphi_{\kappa,\nu}} \sum_{\mu=1}^{T} \mathrm{e}^{\mathrm{j}\vartheta_{\kappa,\mu}} x_\mu(t-\tau_\kappa) \tag{1}$$

where $x_\mu(t)$ is the signal transmitted at antenna μ and $\varphi_{\kappa,\nu}$, $\vartheta_{\kappa,\mu}$, and A_κ are the relative carrier phases at the ν^{th} receive and μ^{th} transmit antenna and the attenuation of the κ^{th} path, respectively. Note that each of the relative carrier phases depends on the distance between the individual antenna element and the scattering object.

Paths whose delays are not sufficiently separated in time cannot be resolved at the receiver. In order to model this effect all such paths are grouped in ascending order of their respective delays into L disjoint sets $\mathcal{D}_\ell$ such that all paths with similar delays form one set. Then, the received signal at the ν^{th} antenna can be decomposed into

$$y_\nu(t) = \sum_{\ell=1}^{L} \sum_{\kappa\in\mathcal{D}_\ell} A_\kappa \mathrm{e}^{\mathrm{j}\varphi_{\kappa,\nu}} \sum_{\mu=1}^{T} \mathrm{e}^{\mathrm{j}\vartheta_{\kappa,\mu}} x_\mu(t-\tau_\kappa). \tag{2}$$

Thus, the propagation coefficient from antenna μ to antenna ν at the delay associated with index ℓ is given as

$$h_{\nu,\mu,\ell} = \sum_{\kappa\in\mathcal{D}_\ell} A_\kappa \mathrm{e}^{\mathrm{j}(\vartheta_{\kappa,\mu}+\varphi_{\kappa,\nu})}. \tag{3}$$

The number of scatterers may vary with delay. Therefore, it is sensible to model this effect by the scatterer count delay profile

$$S_\ell \triangleq |\mathcal{D}_\ell| \tag{4}$$

in addition to the well-known power-delay profile

$$P_\ell \triangleq \frac{1}{R} \sum_{\kappa\in\mathcal{D}_\ell} |A_\kappa|^2. \tag{5}$$

Note that throughout this paper powers are normalized to the number of receive antennas.

The received signal at fixed delay may be written in vector notation as

$$\boldsymbol{y}_\ell = \boldsymbol{H}_\ell \boldsymbol{x} \tag{6}$$

where the entries of the $R \times T$ matrix $\boldsymbol{H}_\ell$ are defined in (3). It is obvious from (3) that those entries show strong statistical dependencies even if A_κ, $\mathrm{e}^{\mathrm{j}\vartheta_{\kappa,\mu}}$, and $\mathrm{e}^{\mathrm{j}\varphi_{\kappa,\nu}}$ are statistically independent for all κ, μ, and ν. These dependencies are examined in greater detail in the following.

Assume without loss of generality that $\mathcal{D}_\ell = \{1, \dots, S_\ell\}$. Define the two $S_\ell \times T$ and $S_\ell \times R$ matrices

$$\boldsymbol{\Theta}_\ell \triangleq \begin{bmatrix} \mathrm{e}^{\mathrm{j}\vartheta_{1,1}} & \dots & \mathrm{e}^{\mathrm{j}\vartheta_{1,T}} \\ \vdots & \ddots & \vdots \\ \mathrm{e}^{\mathrm{j}\vartheta_{S_\ell,1}} & \dots & \mathrm{e}^{\mathrm{j}\vartheta_{S_\ell,T}} \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Phi}_\ell \triangleq \begin{bmatrix} \mathrm{e}^{-\mathrm{j}\varphi_{1,1}} & \dots & \mathrm{e}^{-\mathrm{j}\varphi_{1,R}} \\ \vdots & \ddots & \vdots \\ \mathrm{e}^{-\mathrm{j}\varphi_{S_\ell,1}} & \dots & \mathrm{e}^{-\mathrm{j}\varphi_{S_\ell,R}} \end{bmatrix}, \tag{7}$$

respectively, as well as $\boldsymbol{A}_\ell \triangleq \operatorname{diag}([A_1, \dots, A_{S_\ell}])$. Then, $\boldsymbol{H}_\ell$ may be expressed as

$$\boldsymbol{H}_\ell = \boldsymbol{\Phi}_\ell^{\mathrm{H}} \boldsymbol{A}_\ell \boldsymbol{\Theta}_\ell \tag{8}$$

with $\cdot^{\mathrm{H}}$ denoting the Hermite operator. If there is no line of sight and the antennas within the transmitter array are spaced sufficiently far apart from each other, it is reasonable to assume that the entries within the matrices $\boldsymbol{\Theta}_\ell$ are statistically independent. A corresponding statement holds for the receiver array and the matrices $\boldsymbol{\Phi}_\ell$.

It is not obvious whether the two matrices $\boldsymbol{\Theta}_\ell$ and $\boldsymbol{\Phi}_\ell$ are mutually statistically independent, in practice. In particular, strong dependencies are implied by the single scattering model [6]. In contrast, physical measurements strongly indicate that the correlation between the direction of arrival and the direction of departure are negligible [10, 11]. This supports the proposal that statistical independence of $\boldsymbol{\Theta}_\ell$ and $\boldsymbol{\Phi}_\ell$ is realistic. For the following calculations, independence is assumed.

3 Random Matrix Description

For the further considerations, we condition on the path delay and drop the index ℓ for ease of notation where misunderstandings are unlikely. The performance of communication on a linear channel described by a matrix $\boldsymbol{H}_\ell$ is depends on the R eigenvalues of the covariance matrix $\boldsymbol{H}_\ell \boldsymbol{H}_\ell^{\mathrm{H}}$. In general, not all eigenvalues $\lambda_1, \dots, \lambda_R$ are non–zero, as

$$\operatorname{rank}(\boldsymbol{H}_\ell) \leq \min\{T, R, S_\ell\}. \tag{9}$$

Assuming that the positions of the scattering objects and of the antenna elements within their arrays are random $\boldsymbol{H}_\ell \boldsymbol{H}_\ell^{\mathrm{H}}$ is identified as a random covariance matrix.

The eigenvalues of a large class of random matrices converge in probability to non-random limits as the sizes of the matrices grow over all bounds. The asymptotic limits may serve as good estimates for the eigenvalues in the non–asymptotic case. This has been verified for code–division multiple–access systems in [12] and is likely to generalize to a broader class of communication systems described by large random matrices. In the following, the asymptotic statistics of the eigenvalues are calculated.

Assume that T, R, S_ℓ tend to infinity, but the ratios

$$\zeta \triangleq \frac{T}{S_\ell} \quad \text{and} \quad \xi \triangleq \frac{R}{S_\ell}, \tag{10}$$

remain constant. Then, the distribution of the eigenvalues of

$$\boldsymbol{C}_\ell \triangleq \boldsymbol{H}_\ell \boldsymbol{H}_\ell^{\mathrm{H}} = \boldsymbol{\Phi}_\ell^{\mathrm{H}} \boldsymbol{A}_\ell \boldsymbol{\Theta}_\ell \boldsymbol{\Theta}_\ell^{\mathrm{H}} \boldsymbol{A}_\ell^{\mathrm{H}} \boldsymbol{\Phi}_\ell \tag{11}$$

indeed converges to a non–random limit described by the distribution function [13]

$$\lim_{R\to\infty} \mathrm{F}_{\boldsymbol{C}}(x) \triangleq \lim_{R\to\infty} \frac{1}{R} \left| \{\lambda_i : \lambda_i < x\} \right|. \tag{12}$$

The distribution of the eigenvalues is conveniently represented in terms of its Stieltjes transform

$$G_{C}(s) \triangleq \int \frac{1}{x+s} \, \mathrm{dF}_{C}(x). \tag{13}$$

The Stieltjes transform is not only a tool in random matrix theory that allows for calculation of eigenvalue distributions, but has also a practical meaning in terms of SINRs as discussed later on. In general, the Stieltjes transforms $G_{C_\ell}(\sigma^2)$ provide deep insight into the behavior of multiple–input multiple–output channels in presence of additive white Gaussian noise (AWGN) of variance σ^2.

It follows from [8] that the Stieltjes transform of the asymptotic eigenvalue distribution of the $S_\ell \times S_\ell$ matrix $\boldsymbol{\Phi}_\ell \boldsymbol{\Phi}_\ell^{\mathrm{H}} / S_\ell$ is given by

$$G_{\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{H}}/S}(s) = \frac{1-\xi}{2s} - \frac{1}{2} + \sqrt{\frac{1}{4} + \frac{1+\xi}{2s} + \frac{(1-\xi)^2}{4s^2}}. \tag{14}$$

Note that the same formula applies to the the $S_\ell \times S_\ell$ matrix $\boldsymbol{\Theta}_\ell \boldsymbol{\Theta}_\ell^{\mathrm{H}} / S_\ell$ if ξ is replaced by ζ. It will turn out helpful for calculation of $G_{C_\ell}(s)$ to consider the $S_\ell \times S_\ell$ matrix

$$\tilde{\boldsymbol{C}}_\ell \triangleq \boldsymbol{A}_\ell \boldsymbol{\Theta}_\ell \boldsymbol{\Theta}_\ell^{\mathrm{H}} \boldsymbol{A}_\ell^{\mathrm{H}} \boldsymbol{\Phi}_\ell \boldsymbol{\Phi}_\ell^{\mathrm{H}} \tag{15}$$

instead of the original $R \times R$ matrix (11) in the following. Because of (9), the non–zero eigenvalues of $\boldsymbol{C}_\ell$ are identical to the non–zero eigenvalues of $\tilde{\boldsymbol{C}}_\ell$. Thus, their respective eigenvalue densities $\mathrm{f}(x) \triangleq \frac{\mathrm{d}}{\mathrm{d}x}\mathrm{F}(x)$ satisfy $R\,\mathrm{f}_{C}(x) - R\delta(x) = S_\ell \mathrm{f}_{\tilde{C}}(x) - S_\ell \delta(x)$ which reads in the Stieltjes domain as, cf. (13),

$$R G_{C}(s) - R/s = S_\ell G_{\tilde{C}}(s) - S_\ell / s. \tag{16}$$

Next, we will focus on $G_{\tilde{C}}(s)$ and return to $G_{C}(s)$ later via (16).

3.1 Equal Power Case

The asymptotic eigenvalue distribution depends on the distribution of the powers over the paths with fixed delay. Neglecting shadowing, attenuation is caused by path loss only. This means that all paths with identical delays are attenuated by the same amount giving

$$\boldsymbol{A}_\ell = \sqrt{\frac{P_\ell}{R S_\ell}} \mathbf{I}. \tag{17}$$

with (5). Note that this model includes multi–path Rayleigh fading. Moreover, it is even more accurate than the Rayleigh fading model, as it does not need to approximate the sum of the complex phases in (3) by a complex Gaussian random variable.

The equal power assumption gives

$$\tilde{\boldsymbol{C}}_\ell \triangleq \frac{P_\ell}{\xi} \frac{\boldsymbol{\Theta}_\ell \boldsymbol{\Theta}_\ell^{\mathrm{H}}}{S_\ell} \frac{\boldsymbol{\Phi}_\ell \boldsymbol{\Phi}_\ell^{\mathrm{H}}}{S_\ell}. \tag{18}$$

Obviously, $\tilde{\boldsymbol{C}}_\ell$ is the product of two independent random matrices. The Stieltjes transforms corresponding to the factors in the product are available from (14).

A machinery that yields the Stieltjes transform corresponding to a matrix product given the Stieltjes transforms corresponding to the factors is found in [13, Sec. 3.6] and called *multiplicative free convolution.* Its key tool is the S–transform $S(z)$ which is defined in terms of Stieltjes transforms as

$$S(z) \triangleq \frac{z+1}{z}\Upsilon^{-1}(z) \qquad \text{with} \qquad \Upsilon(s) \triangleq \frac{G\left(-s^{-1}\right)}{-s} - 1 \quad \text{and} \quad \left(\Upsilon^{-1}\circ\Upsilon\right)(s) = s. \tag{19}$$

The S–transforms corresponding to the factors in (18) can be shown with (19) to be

$$S_{\boldsymbol{\Theta\Theta}^{\mathrm{H}}/S}(z) = \frac{1}{z+\zeta} \qquad \text{and} \qquad S_{\boldsymbol{\Phi\Phi}^{\mathrm{H}}/S}(z) = \frac{1}{z+\xi}, \tag{20}$$

respectively. Since the S–transform of a product is the product of the factors' S–transforms [13], we get

$$S_{\frac{\xi}{P}\tilde{\boldsymbol{C}}}(z) = \frac{1}{(z+\zeta)(z+\xi)}. \tag{21}$$

Definitions (19) yield

$$s^2 G^3_{\frac{\xi}{P}\tilde{\boldsymbol{C}}}(s) + s(\zeta+\xi-2)G^2_{\frac{\xi}{P}\tilde{\boldsymbol{C}}}(s) + \big(s + (\zeta-1)(\xi-1)\big)G_{\frac{\xi}{P}\tilde{\boldsymbol{C}}}(s) - 1 = 0. \tag{22}$$

Note that $\mathrm{f}_{\frac{\xi}{P}\boldsymbol{C}}(x) = \frac{P}{\xi}\mathrm{f}_{\boldsymbol{C}}(\frac{P}{\xi}x)$ which reads in the Stieltjes domain as $G_{\frac{\xi}{P}\boldsymbol{C}}(s) = \frac{P}{\xi}G_{\boldsymbol{C}}(\frac{P}{\xi}s)$. Together with (22), this yields via (16)

$$s^2\xi G^3_{\boldsymbol{C}}(s) + s\left(1+\zeta-2\xi\right)G^2_{\boldsymbol{C}}(s) + \left(\frac{s}{P_\ell} + \zeta/\xi - 1 + \xi - \zeta\right)G_{\boldsymbol{C}}(s) - \frac{1}{P_\ell} = 0. \tag{23}$$

Although (23) can be resolved with respect to $G_{\boldsymbol{C}}(s)$ via Cardano's formula, the result is omitted here, as it complicated and lacks of new insight.

Additional insight into (23) is obtained defining the ratios

$$\beta \triangleq \frac{T}{R} = \frac{\zeta}{\xi} \qquad \text{and} \qquad \rho \triangleq \frac{S_\ell}{R} = \frac{1}{\xi} \tag{24}$$

Note that in theory of code–division multiple–access (CDMA) [14] where the numbers of receive and transmit antennas correspond to the spreading factor and the number of users, respectively, β is known as the *load* of the system. For convenience, it is named system load in the context of antenna arrays, too. With (24), the parameters ζ and ξ may be eliminated from (23), giving:

$$s^2 P_\ell G^3_{\boldsymbol{C}}(s) + sP_\ell(\rho+\beta-2)G^2_{\boldsymbol{C}}(s) + \big(s\rho + P_\ell(\rho-1)(\beta-1)\big)G_{\boldsymbol{C}}(s) - \rho = 0. \tag{25}$$

In (25), the Stieltjes transform is parameterized by the system load β giving the size ratio of the channel matrix and the *richness* of the channel ρ which is the number of scatterers normalized to the number of receive antennas. This richness in scattering is linked in literature to high channel capacities for communication through antenna arrays.

Previously, the asymptotic eigenvalue distribution was characterized in terms of its Stieltjes transform. In the following, examples are given that demonstrate how to use the Stieltjes transform to derive more intuitive performance measures like SINRs and channel capacity.

Optimum Linear Detection: For unique path delay, i.e. $L = 1$, (25), can be easily used to express the SINR with optimum linear detection (Wiener filtering, optimum linear beam forming at receiver site) making use of standard results in mean–square estimation, e.g. [8, Eq. (27)],

$$\beta \frac{SINR}{1 + SINR} = 1 - \sigma^2 G_{\boldsymbol{C}}(\sigma^2). \tag{26}$$

Fig. 1 shows that rich scattering is crucial for high SINR. The importance of rich scattering has been conjectured in various references. In this work, it is found for the first time by analysis of an abstract channel model which confirms the sensibility of the proposed random matrix approach. In addition, we find that rich scattering is the more important the higher the system load.

The special case of independent entries in $\boldsymbol{H}$ is included in (25) and appears for $\rho \to \infty$, i.e. the relative number of scatterers goes to infinity. As expected, in that case $G_{\boldsymbol{C}}(s)$ in (25) becomes equivalent to $G_{\boldsymbol{\Theta}\boldsymbol{\Theta}^{\rm H}/S}(s)$ in (14) when ξ is replaced by β.

Nulling: Assume the receiver does not apply the optimum linear beam former, but simply processes the received signal in such a way that crosstalk is completely nulled out. In matrix notation, this type of processing is identified as channel inversion. In CDMA literature [14], it is known as decorrelation.

For unique path delay the SINR is given as

$$SINR' = \frac{1}{\dfrac{\sigma^2}{T} \displaystyle\sum_{i=1}^{T} \frac{1}{\lambda'_i}} \tag{27}$$

where $\lambda'_i, 1 \leq i \leq T$ are the eigenvalues of $\boldsymbol{C}' \triangleq \boldsymbol{H}^{\rm H}\boldsymbol{H}$. Thus, the SINR for large scale systems is given as $1/(\sigma^2 G_{\boldsymbol{H}^{\rm H}\boldsymbol{H}}(0))$.

The Stieltjes transform of $\boldsymbol{H}^{\rm H}\boldsymbol{H}$ can be obtained from the Stieltjes transform of $\boldsymbol{C} = \boldsymbol{H}\boldsymbol{H}^{\rm H}$ given in (25) with the help of (16) if S_ℓ and $\tilde{\boldsymbol{C}}$ in (16) are replaced by T and $\boldsymbol{C}'$, respectively. After some elementary algebra and rewriting in terms of ζ instead of the richness, this yields

$$s^2 P_\ell \beta \zeta G^3_{\boldsymbol{C}'}(s) + s P_\ell (\beta + \zeta - 2\beta\zeta) G^2_{\boldsymbol{C}'}(s) + \big(s + P_\ell(\beta - 1)(\zeta - 1)\big) G_{\boldsymbol{C}'}(s) - 1 = 0. \tag{28}$$

Setting $s = 0$ immediately gives the asymptotic SINR as

$$SINR' = \frac{P_\ell}{\sigma^2}(1 - \beta)(1 - \zeta). \tag{29}$$

Obviously, the parameter ζ has similar meaning as the system load β. With respect to this symmetry, we call it *channel load*, in the following. In order to ensure full rank of $\boldsymbol{C}'$, both system and channel load are required to be smaller than one. This means that the richness has to be larger than the system load.

The symmetry in ζ and β also extends to the SINR of optimum beam forming. Plugging (26) into (25) gives

$$\frac{\sigma^2}{P_\ell} SINR^3 + \left(2\frac{\sigma^2}{P_\ell} - (\beta - 1)(\zeta - 1)\right) SINR^2 + \left(\frac{\sigma^2}{P_\ell} + \beta + \zeta - 2\right) SINR - 1 = 0. \tag{30}$$

The degradation of the nulling strategy compared to (30) keeps small for $\max\{\beta, \zeta\} \ll 1$.

Channel Capacity: We assume that only the receive array is aware of the channel. Therefore, the transmit array cannot adapt the statistics of its signals to the channel. With this preliminary, channel capacity per receive antenna in presence of complex

AWGN is given as $C(\sigma^2) = \sum_{i=1}^{T} \log\left(1 + \lambda_i'/\sigma^2\right)/R$ with eigenvalues λ_i' defined as in (27). For asymptotically large arrays, this means

$$C(s) = \beta \int \log\left(1 + \frac{x}{s}\right) \mathrm{dF}_{C'}(x). \qquad \Rightarrow \qquad \frac{\partial C}{\partial s} = \beta G_{C'}(s) - \frac{\beta}{s}. \tag{31}$$

This can be plugged into (28) and expressed in terms of system load and richness giving the following differential equation for the channel capacity

$$\left(s\frac{\partial C}{\partial s}\right)^3 + (\beta + \rho + 1)\left(s\frac{\partial C}{\partial s}\right)^2 + \left(\frac{s\rho}{P_\ell} + \rho + \beta\rho + \beta\right)\left(s\frac{\partial C}{\partial s}\right) + \rho\beta = 0 \tag{32}$$

with the boundary condition $\lim_{s\to\infty} C(s) = 0$. The numerical solution to this differential equation is depicted in Fig. 2. For rich scattering, the capacity grows almost linearly with increasing number of transmit antennas until unit system load is reached. For higher system loads, no free dimensions are left in the signal space and capacity grows only logarithmically through increase in transmitted signal power. For poor scattering, the capacity of the antenna array degrades severely.

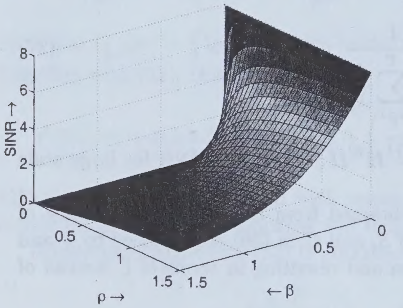


Figure 1: SINR as function of system load and richness for $10\log_{10}(P_\ell/\sigma^2) = 9$ dB.

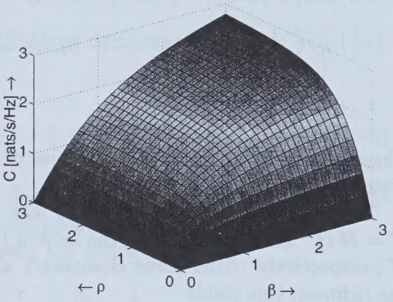


Figure 2: Channel capacity of the antenna array per unit number of receive antennas with $10\log_{10}(P_\ell/\sigma^2) = 9$ dB.

Moments of the Eigenvalues: For some applications, e.g. weighted linear interference cancellation [15, 16], the moments of the eigenvalue density are of primary importance. In terms of binomial coefficients, they can be shown to be

$$m_n' \triangleq \int x^n \mathrm{dF}_{C'}(x) = P_\ell^n \sum_{i=0}^{n-1}\sum_{j=0}^{n-1} \binom{n}{i}\binom{n}{j}\binom{n}{i+j+1} \frac{\beta^i \zeta^j}{n} \qquad \forall\, n \geq 0. \tag{33}$$

The moments of $\boldsymbol{C}$ can be easily derived from the moments of $\boldsymbol{C}'$ and are given by $\beta m_n'$ as the non–zero eigenvalues are identical for the two matrices and the zero eigenvalues do not contribute to the moments for $n > 0$. The correction factor β results from the different sizes of the respective matrices.

3.2 Multiple Delays

The previous random matrix results have shown that the eigenmodes of the channel conditioned on a fixed path delay are determined by only two parameters: the richness

ρ and the system load β. The latter is rather a system parameter than a property of the channel. Therefore, the *physical* channel conditions are canonically determined by the richness ρ. Certainly, the richness may depend on the path delay. Instead of the scatterer–count delay profile introduced in (4) we can use the richness delay profile ρ_ℓ to characterize the channel and address the question, how the richness delay profile affects the space–time nature of the channel.

Stack the symbol vectors $\boldsymbol{x}[k]$ transmitted at subsequent time–instances k one below each other into a single vector of transmitted data. Without loss of conceptual scope, the transmission in space and time can be written as a single matrix equation:

$$\begin{bmatrix} \vdots \\ \boldsymbol{y}[0] \\ \boldsymbol{y}[1] \\ \vdots \end{bmatrix} = \underbrace{\begin{bmatrix} \ddots & \ddots & & \ddots & \ddots & \ddots & & \\ \cdots & 0 & \boldsymbol{H}_L & \cdots & \boldsymbol{H}_1 & \boldsymbol{H}_0 & 0 & \cdots & \\ & \cdots & 0 & \boldsymbol{H}_L & \cdots & \boldsymbol{H}_1 & \boldsymbol{H}_0 & 0 & \cdots \\ & & \ddots & \ddots & & \ddots & \ddots & \ddots & \ddots \end{bmatrix}}_{\triangleq \mathcal{H}} \begin{bmatrix} \vdots \\ \boldsymbol{x}[0] \\ \boldsymbol{x}[1] \\ \vdots \end{bmatrix} \tag{34}$$

Note that the space–time channel matrix $\mathcal{H}$ is a circulant matrix in the space of $\infty \times \infty$ matrices on $\mathbb{C}^{R\times T}$. Therefore, it can be decomposed into

$$\mathcal{H} = \mathcal{T}_R \mathcal{L} \mathcal{T}_T^{\mathrm{H}} \tag{35}$$

where $\mathcal{T}_i$ are $\infty \times \infty$ Fourier matrices on the set of $i \times i$ identity matrices that are scaled by a complex scalar and $\mathcal{L}$ is an $\infty \times \infty$ diagonal matrix on the set of $R \times T$ matrices. The Fourier matrices are unitary with respect to both the matrix entries as well as the underlying complex scalar entries. Thus, the singular values of the space–time channel matrix are identical to those of $\mathcal{L}$.

Since the space–time channel matrix is circulant, the non–zero matrix–valued entries of $\mathcal{L}$ are given as the Fourier transform of the first row of $\mathcal{H}$ in $\mathbb{C}^{R\times T}$ [20]. Therefore, the i^{th} matrix–valued diagonal element can be written as

$$\breve{\boldsymbol{H}}_i = \sum_{\ell=1}^{L} \boldsymbol{H}_\ell \, \mathrm{e}^{\mathrm{j}\alpha_{i,\ell}} \tag{36}$$

with some $\alpha_{i,\ell}$ being determined by the Fourier transform. The complex phase factors $\exp(\mathrm{j}\alpha_{i,\ell})$ do not change the statistics of the random matrices $\boldsymbol{H}_\ell$, as the entries of $\boldsymbol{H}_\ell$ are assumed to be circularly symmetric, cf. (3). Therefore, $\alpha_{i,\ell}$ can be assumed to be zero without changing the expected eigenvalue distribution of $\breve{\boldsymbol{H}}_i \breve{\boldsymbol{H}}_i^{\mathrm{H}}$.

The space–time channel matrix $\mathcal{H}$ is block–diagonal. Therefore, its expected singular value distribution is the mean of the expected singular value distributions of the block matrices $\breve{\boldsymbol{H}}_i$. The expected singular value distributions of all the block matrices are identical, as $\alpha_{i,\ell}$ may be assumed zero without changing the expected singular value distribution of the block matrices. Therefore, the expected singular value distributions of

$$\mathcal{H} \qquad \text{and} \qquad \sum_{\ell=1}^{L} \boldsymbol{H}_\ell \tag{37}$$

are identical. Note that this holds regardless of the number of receive and transmit antennas. In the asymptotic case, however, it does not only hold for the expected singular value distributions, but for the *limiting* singular value distribution. This limiting distribution does not depend on the explicit realizations of the matrices $\boldsymbol{H}_\ell$.

The summation over the channel matrices conditioned on the delay can be written with (8) as an inner product of composite matrices

$$\sum_{\ell=1}^{L} \boldsymbol{H}_\ell = \underbrace{\left[\boldsymbol{\Phi}_1^{\mathrm{H}}\, \boldsymbol{\Phi}_2^{\mathrm{H}}\, \ldots\, \boldsymbol{\Phi}_L^{\mathrm{H}}\right]}_{\triangleq\, \boldsymbol{\Phi}^{\mathrm{H}}} \underbrace{\operatorname{diag}(\boldsymbol{A}_1, \ldots, \boldsymbol{A}_L)}_{\triangleq\, \boldsymbol{A}} \underbrace{\left[\boldsymbol{\Theta}_1^{\mathrm{H}}\, \boldsymbol{\Theta}_2^{\mathrm{H}}\, \ldots\, \boldsymbol{\Theta}_L^{\mathrm{H}}\right]^{\mathrm{H}}}_{\triangleq\, \boldsymbol{\Theta}}. \tag{38}$$

Note that (38) describing the channel with multiple delays takes the same form as (8) which is conditioned on a particular path delay and the sum in (38) just adds together all propagation paths regardless of their delays and stacks them in the composite matrices $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$. The dimensions of the composite matrices are $R \times S_{\max}$ and $T \times S_{\max}$, respectively, with

$$S_{\max} = \sum_{\ell=1}^{L} S_\ell. \tag{39}$$

Therefore, a channel with varying delays is equivalent to a channel with identical delays where the richness per delay is replaced by the *total richness*

$$\rho \triangleq \sum_{\ell=1}^{L} \rho_\ell \tag{40}$$

while the system load remains unaffected. This means that for large antenna arrays *the eigenvalues of the space–time channel matrix do not change if delay times of particular paths vary.* Moreover, there is no need to distinguish between the distributions of path attenuations conditioned on different delays. Therefore, neither the scatterer–count nor the power–delay profile are necessary to define the asymptotic eigenvalue distribution of the space–time channel matrix $\boldsymbol{\mathcal{H}}$. It is sufficiently characterized by

- the system load β,
- the total richness ρ,
- and the distribution of the attenuations $A_\kappa, 1 \leq \kappa \leq S_{\max}$.

If all attenuations are identical, the results found in Section 3.1 by conditioning on a particular delay apply to the general space–time channel with multi–path propagation in both space and time.

For systems with finite number of antennas the *expected* eigenvalue distribution was found unaffected by inter–symbol interference due to multiple delays. However, the averaging in (38) drives the fluctuations of the empirical distributions to decrease with the number of resolvable path delays L.

4 Summary and Conclusions

A random matrix model for communication through antenna arrays has been introduced. The relative number of scatterers (richness) in the channel was found to be a key parameter to understand the behavior of channel's eigenmodes and therefore its capacity. The problem has been found analytically tractable, and the Stieltjes transforms and the moments of the eigenvalue densities were calculated. Inter–symbol interference has been found beneficial for arrays with finite number of antennas, as it reduces fluctuations of the eigenvalues, while it has no effect on the eigenvalues of the space–time channel matrix for asymptotically large arrays.

Acknowledgment

The author would like to thank C. Mecklenbräuker, M. Toeltsch, E. Bonek, A. Molisch, and H. Hofstetter for helpful discussions.

References

[1] G. Foschini. Layered space–time architecture for wireless communication in a fading environment when using multi–element antennas. *Bell Labs Techn. J.*, 1(2):41–59, 1996.

[2] G. Foschini and M. Gans. On limits of wireless communications in a fading environment when using multiple antennas. *Wireless Personal Commun.*, 6:311–335, 1998.

[3] G. Raleigh and J. Cioffi. Spatio–temporal coding for wireless communication. *IEEE Trans. Commun.*, 46(3):357–366, Mar. 1998.

[4] V. Tarokh, N. Seshadri, and A. Calderbank. Space–time codes for high data rate wireless communication: Performance criterion and code construction. *IEEE Trans. Inform. Th.*, 44(2):744–765, Mar. 1998.

[5] H. Bölcskei, D. Gesbert, and A. Paulraj. On the capacity of wireless systems employing OFDM–based spatial multiplexing. Submitted to *IEEE Trans. Commun.*

[6] J. Liberti and T. Rappaport. *Smart Antennas for Wireless Communications.* Prentice–Hall, 1999.

[7] E. Telatar. Capacity of multi–antenna Gaussian channels. Bell–Labs internal report, unpublished, June 1995.

[8] D. Tse and S. Hanly. Linear multiuser receivers: Effective interference, effective bandwidth and user capacity. *IEEE Trans. Inform. Th.*, 45(2):641–657, Mar. 1999.

[9] J. Andersen. Antenna arrays in mobile communications: Gain, diversity, and channel capacity. *IEEE Ant. & Propag. Mag.*, 42(2):12–16, Apr. 2000.

[10] M. Steinbauer, D. Hampicke, G. Sommerkorn, A. Schneider, A. Molisch, R. Thomä, and E. Bonek. Array measurement of the double–directional mobile radio channel. In *Proc. of IEEE VTC*, Tokio, Japan, May 2000.

[11] M. Steinbauer, A. Molisch, and E. Bonek. The double–directional mobile radio channel. Submitted to *IEEE Ant. & Propag. Mag.*

[12] R. Müller. Multiuser receivers for randomly spread signals: Fundamental limits with and without decision–feedback. To appear in *IEEE Trans. Inform. Theory.*

[13] D. Voiculescu, K. Dykema, and A. Nica. *Free Random Variables.* American Mathem. Soc., Providence, RI, 1992.

[14] S. Verdú. *Multiuser Detection.* Cambridge University Press, New York, 1998.

[15] R. Müller and S. Verdú. Spectral efficiency of low–complexity multiuser detection. In *Proc. of IEEE ISIT*, p. 439, Sorrento, Italy, June 2000.

[16] D. Guo, L. Rasmussen, and T. Lim. Linear parallel interference cancellation in long–code CDMA multiuser detection. *IEEE J. Sel. A. Commun.*, 17(12):2074–2081, Dec. 1999.

[17] S. Hanly and D. Tse. Resource pooling and effective bandwidth in CDMA networks with multiuser receivers and spatial diversity. Submitted to *IEEE Trans. Inform. Theory.*

[18] J. Proakis. *Digital Communications.* McGraw–Hill, New York, 3rd ed., 1995.

[19] M. Töltsch, J. Laurila, K. Kalliola, A. Molisch, and E. Bonek. Wideband 3–D mobile radio channel measurements in urban environment. Submit. to *IEEE Ant. & Propag. Mag.*

[20] P. Davis. *Circulant Matrices.* John Wiley & Sons, New York, 1979.

Biomolecular Conformations as Metastable Sets of Markov Chains

Ch. Schütte and W. Huisinga[†]

Institute for Mathematics I, Free University Berlin, Arnimallee 2-6, 14195 Berlin
schuette, huisinga@math.fu-berlin.de

Abstract

The function of many important biomolecules comes from their dynamic properties and their ability to make statistically rare switches between different *conformations.* Recent investigations demonstrated that (a) these conformations can be understood as *metastable* or *almost invariant* sets of certain Markov chains related with the dynamical behavior of the molecular system and that (b) these sets can efficiently be computed via eigenvectors of some associated transfer operators. The article presents in general the idea of computing metastable sets of Markov chains via the dominant eigenmodes of some associated transfer operator, refers its rigorous justification for specific chains, and illustrates the application of the resulting algorithm to problems arising from molecular systems.

1 Introduction

The chemically interesting *function* of many important biomolecules, like proteins or enzymes, results from their *dynamical* properties, particularly from their ability to undergo so-called *conformational transitions* (cf. [22]). In a conformation, the large scale geometric structure of the molecule is understood to be conserved, whereas on smaller scales the system may well rotate, oscillate or fluctuate. Recently, DEUFLHARD ET AL. demonstrated that conformations can be understood as *almost invariant* or *meta-stable* sets of the Hamiltonian system governing the molecular dynamics [5]. In 1996, DELLNITZ and JUNGE demonstrated that such almost invariant sets of discrete dynamical systems with small random perturbations can be identified numerically via certain "dominant" eigenvectors of the Markov operator associated with the perturbed system [4].

By transferring this idea to statistical mechanics, SCHÜTTE et al. showed that the almost invariant sets of dynamical fluctuations in statistical molecular ensembles can be determined via the "dominant" eigenvectors of a specific class of Markov operators associated with discrete time Markov chains with nonlinear state space [16, 17]. It has been demonstrated that, even for larger (bio)molecules, the eigenvectors of interest can be computed efficiently and allow to identify the desired almost invariant sets [6, 16]. This allows for the first time to identify dynamical conformations of molecular ensembles including their stability life spans and the rate of transitions between them [10].

[†] Supported by Deutsche Forschungsgesellschaft, Germany

The present article summarizes the basic ideas by reformulating it in a rather general framework for a large class of Markov chains. The presentation is tailored to demonstrate how the molecular applications and such of DELLNITZ ET AL. result from the general framework (s. Sec. 5). The article closes with a specific example which illustrates that the suggested approach allows to uncover purely dynamically induced metastabilities which are *not* explained by the common physical beliefs (s. Sec. 6).

2 Markov Chains and Transfer Operators

Consider a probability space $(\mathbf{X}, \mathcal{A}, \mu)$, where $\mathbf{X} \subset \mathbf{R}^m$ for some $m \in \mathbf{N}$ denotes the state space, $\mathcal{A}$ the Borel σ–algebra on $\mathbf{X}$ and μ a probability measure on $\mathcal{A}$. By $t \in \mathbf{T}$ we denote the discrete or continuous time, i.e., $\mathbf{T} = \mathbf{R}_0^+$ or $\mathbf{T} = \mathbf{N}$.

A function $p : \mathbf{T} \times \mathbf{X} \times \mathcal{A} \to [0,1]$ is called a *stochastic transition kernel* [3, 12], if (i) $p(t, x, \cdot)$ is a probability measure on $\mathcal{A}$ for every $t \in \mathbf{T}$, $x \in \mathbf{X}$ and furthermore, $p(0, x, \mathbf{X} \setminus \{x\}) = 0$ for every $x \in \mathbf{X}$, (ii) $p(t, \cdot, A)$ is measurable for every $t \in \mathbf{T}$, $A \in \mathcal{A}$, and (iii) $p(\cdot, x, A)$ satisfies the Chapman–Kolmogorov equation [7, 12]

$$p(t+s, x, A) \;=\; \int_{\mathbf{X}} p(t, x, \mathrm{d}y)\, p(s, y, A) \tag{1}$$

for all $t, s \in \mathbf{T}$, $x \in \mathbf{X}$ and $A \in \mathcal{A}$. The family $\{X_t\}_{t \in \mathbf{T}}$ is called a homogeneous Markov process, if [3, 12]

$$\mathbf{P}[X_{t+s} \in A \,|X_s = x] \;=\; p(t, x, A) \tag{2}$$

for all $s, t \in \mathbf{T}$, $x \in \mathbf{X}$ and and $A \in \mathcal{A}$. Thus $p(t, x, C)$ is the probability that the Markov process started in x stays in A after the time span t.

In order to determine the essential statistical behavior of the Markov process, we have to be able to analyze the evolution of the process for some given statistical distribution of initial states. This is usually done by considering initial probability distributions rather than initial states for starting the Markov process. Assume that X_0 is distributed according to some probability measure ν, i.e., $\mathbf{P}[X_0 \in C] = \nu(C)$. Then, the evolution of the distribution of X_t is given by

$$\mathbf{P}_\nu[X_t \in C] \;=\; \int_X \nu(\mathrm{d}x) p(t, x, C).$$

Of special interest are probability measures that are invariant w.r.t. X_t, i.e., which satisfy $\mathbf{P}_\mu[X_t \in C] = \mu(C)$ for every $t \in \mathbf{T}$; they are also called stationary probability distributions.

Governed by application to biomolecular systems, we focus our attention to Markov processes that exhibit a unique stationary probability distribution, denoted by μ. Furthermore, we are interested in determine the internal fluctuations within the stationary distribution, which then will enable us to identify metastable subsets of the state space. As we will see in the next section, the description of such internal fluctuation requires to consider only initial probability distribution that are absolutely continuous w.r.t. the stationary distribution μ. The evolution of some ensemble of initial states distributed according to $\nu(\mathrm{d}x) = u(x)\mu(\mathrm{d}x)$ for some density $u \in L^1(\mu)$ may be described by the *propagator* or *forward transfer operator*

$$P_t u(y)\,\mu(\mathrm{d}y) \;=\; \int_{\mathbf{X}} p(t,x,\mathrm{d}y)\,u(x)\,\mu(\mathrm{d}x),$$

which is well–defined on $L^1(\mu)$ [13, Chpt. 4]. Its adjoint operator, the *backward transfer operator* acts on $L^\infty(\mu)$ according to [7]

$$T_t u(x) \;=\; \mathbf{E}_x[u(X_t)] = \int_{\mathbf{X}} u(y)\,p(t,x,\mathrm{d}y), \tag{3}$$

where $\mathbf{E}_x[u(X_t)]$ denotes the expectation of an observable $u : \mathbf{X} \to \mathbf{C}$ under the condition that the process $\{X_t\}$ has been started at $t = 0$ in x. Hence, in terms of the duality bracket $\langle\cdot,\cdot\rangle_\mu$ between $L^\infty(\mu)$ and $L^1(\mu)$ we have $\langle T_t u, v\rangle_\mu = \langle u, P_t v\rangle_\mu$ for $u \in L^\infty(\mu)$ and $v \in L^1(\mu)$.

Since $p(t,\cdot,\cdot)$ is a transition kernel, the thereby defined operator P_t is a *Markov operator* on $L^1(\mu)$. Furthermore, the semigroup property of the Markov process implies that $\{P_t\}_{t\in\mathbf{T}}$ is a semigroup of Markov operators. Due to the properties of the transition kernel and the definition of the backward transfer operator, we have for every $t \in \mathbf{T}$, $T_t\mathbf{1}_{\mathbf{X}} = \mathbf{1}_{\mathbf{X}}$, where $\mathbf{1}_A$ denotes the characteristic function of the subset A. Note, that $\mathbf{1}_X$ represents the stationary probability distribution μ. The above equality does in general not hold for the forward transfer operator, because P_t depends on the probability measure μ. However, if we assume μ to be invariant, we also get $P_t\mathbf{1}_{\mathbf{X}} = \mathbf{1}_{\mathbf{X}}$ for all $t \in \mathbf{T}$. In other words, $\mathbf{1}_{\mathbf{X}}$ is an invariant density of P_t, whenever μ is invariant.

3 Transition Probabilities and Almost Invariant Sets

Assume that the Markov process is initially distributed according to μ. The *transition probability* $p(s,C,D)$ within the stationary distribution from $C \in \mathcal{A}$ to $D \in \mathcal{A}$ within the time span s is defined as the conditional probability

$$p(s,C,D) \;=\; \mathbf{P}_\mu[X_s \in D\,|X_0 \in C]. \tag{4}$$

The similar symbols for both the transition probability $p(s,C,D)$ and for the transition kernel $p(s,x,C)$ corresponding to the process emphasizes the strong relation to (2), which, in addition to the above assumption, allows to rewrite the transition probability as

$$p(s,C,D) \;=\; \frac{1}{\mu(C)}\int_C p(s,x,D)\,\mu(\mathrm{d}x). \tag{5}$$

The transition probabilities quantify the *dynamical fluctuations within the stationary distribution.* Using the duality bracket $\langle\cdot,\cdot\rangle_\mu$, the definitions of the transfer operators T_t and P_t yield

$$p(s,C,D) = \frac{\langle T_s\mathbf{1}_D, \mathbf{1}_C\rangle_\mu}{\langle \mathbf{1}_C, \mathbf{1}_C\rangle_\mu} = \frac{\langle \mathbf{1}_D, P_s\mathbf{1}_C\rangle_\mu}{\langle \mathbf{1}_C, \mathbf{1}_C\rangle_\mu}. \tag{6}$$

Following [4], we call some subset $C \in \mathcal{A}$ *almost invariant*, whenever the fraction of systems within the distribution that stay in C after some characteristic time span $s \in \mathbf{T}$ is close to 1:

$$C \text{ almost invariant} \iff p(s,C,C) \approx 1.$$

Almost invariance may equivalently be characterized by $p(s,C,\mathbf{X}\setminus C)\approx 0$, which allows to relate it to the semigroup of forward transfer operators $\{P_t\}_{t\in\mathbf{T}}$ by the following general identity [21]:

$$\left\| P_s \frac{1}{\mu(C)} \mathbf{1}_C - \frac{1}{\mu(C)} \mathbf{1}_C \right\|_1 = 2\, p(s,C,\mathbf{X}\setminus C). \tag{7}$$

4 Identification of Almost Invariant Sets

Since P_s is a Markov operator its $L^1(\mu)$–spectrum is contained in the unit ball $\{\lambda \in \mathbf{C} : |\lambda| \le 1\}$. Every *invariant* density $u \in L^1(\mu)$ of P_s satisfies $P_s u = u$ and therefore is an eigenvector of P_s corresponding to the eigenvalue $\lambda = 1$, the so–called *Perron root.* Since μ is assumed to be invariant, in particular $u = \mathbf{1}_X$ is an invariant density.

Whenever a proper subset C of $\mathbf{X}$ is invariant under the Markov process, i.e., $p(s,x,\mathbf{X}\setminus C) = 0$ for all $x \in C$, the density $u = \mathbf{1}_C/\mu(C)$ is an eigenvector corresponding to $\lambda = 1$. Due to our above characterization, the set $C \in \mathcal{A}$ is almost invariant if $p(s,C,\mathbf{X}\setminus C) \approx 0$, which via formula (7) implies that $\mathbf{1}_C/\mu(C)$ is an approximate invariant density, i.e., an approximate normalized eigenvector associated with an eigenvalue close to the Perron root $\lambda = 1$. This motivates the following *algorithmic strategy*:

> *Invariant* sets can be identified via eigenvectors of P_s corresponding to the Perron root $\lambda = 1$, while *almost invariant* sets may be identified via eigenvectors corresponding to eigenvalues $|\lambda| < 1$ close to the Perron root $\lambda = 1$.

This strategy has first been proposed by DELLNITZ AND JUNGE [4] for discrete dynamical systems with weak random perturbations and has been successfully applied to molecular dynamics in different contexts [16, 17, 18]; it is justified in more detail in [19]. For a detailed description of the resulting *identification algorithm* in the molecular context, cf. [6, 19]; for an illustration of its basic idea see Fig. 1.

We want to emphasize that almost invariance is defined herein with respect to some pre–selected invariant probability measure μ that describes the stationary ensemble under consideration. The above algorithmic strategy requires uniqueness of the invariant measure. For its numerical realization via an eigenvalue problem we moreover need that the remaining spectrum of P_s is strictly bounded away from the Perron root, i.e., $\lambda = 1$ must be an isolated, simple eigenvalue of P_s. Additionally, the physical interpretation of the ensemble excludes other eigenvalues than $\lambda = 1$ *on* the unit circle or, equivalently, we exclude asymptotic periodicity of P_s.

We introduce the following two *fundamental conditions* on the forward transfer operator P_s that are sufficient to guarantee the desired properties:

(C1) P_s is asymptotically stable, i.e., $(P_s)^n u \to \mathbf{1}_\mathbf{X}$ in $L^1(\mu)$ for every density $u \in L^1(\mu)$ as $n \to \infty$.

(C2) The essential spectrum of P_s in $L^1(\mu)$—and sometimes in $L^2(\mu)$—is strictly bounded away from $|\lambda| = 1$.

For details about the essential spectral radius (and its relation to asymptotic properties of transfer operators) see [9].

5 Examples from Molecular Dynamics (MD)

Stochastically perturbed Hamiltonian dynamics. Classical models for molecular motion describe the molecular system via coupled equations of motion for the N atoms in the system. For a differentiable potential function $V : X \subset \mathbf{R}^d \to \mathbf{R}$, the *Hamiltonian equations of motion* are given by

$$\dot{q} = M^{-1}p, \qquad \dot{p} = -\nabla_q V(q) \tag{8}$$

where $d = 3N$, and q and p denote the positions and the momenta of the atoms in the system, respectively and M the diagonal mass matrix [8]. Let $q_t = q_t(q_0, p_0), p_t = p_t(q_0, p_0)$ denote the solution of (8) for the initial data q_0, p_0. It is well–known that the *canonical measure*

$$\mu_{\text{can}}(\mathrm{d}q\,\mathrm{d}p) = \underbrace{\frac{1}{Z_q}\exp(-\beta V(q))\mathrm{d}q}_{\mu_Q(\mathrm{d}q)} + \underbrace{\frac{1}{Z_p}\exp(-\frac{\beta}{2}p^t M^{-1}p)\mathrm{d}p}_{\mu_P(\mathrm{d}p)}$$

corresponding to the so–called inverse temperature β is invariant w.r.t. the evolution process of (8). Since μ_{can} is known to decribe molecular ensembles with constant temperature, as they appear in biomolecular applications, this is the stationary distribution of interest. In the following, we restrict our attention to potential functions V that allow to normalize μ_{can} to a probability measure.

Aiming at the identification of molecular conformations, we choose some *fixed* observation time span $\tau > 0$ and introduce the *Hamiltonian stochastic system*

$$Q_{n+1} = q_\tau(Q_n, P_n), \quad n = 1, 2, \ldots, \tag{9}$$

where $\{P_n\}_{n\in\mathbf{N}}$ is an i.i.d. sequence of random variables, each distributed according to μ_P, i.e., $\mathbf{P}[P_n \in A] = \mu_P(A)$ [16, 17]. Metastable sets of this Markov chain are *biomolecular conformations* of ensembles of molecular systems with constant temperature [17]. The stochastic transition function corresponding to (9) is given by

$$p(q, A) = \int_{\mathbf{R}^d} \mathbf{1}_A(q_\tau(q,p))\; \mu_P(\mathrm{d}p)$$

for all $A \in \mathcal{B}(X)$; in [16] it is shown that μ_Q is stationary for the process. Exploiting properties of the Hamiltonian equation of motion, the corresponding propagator $P : L^1(\mu_Q) \to L^1(\mu_Q)$ may be written as

$$Pv(y) = \int_{\mathbf{R}^d} v\,(q_\tau(q,p))\; \mu_P(\mathrm{d}p)$$

for $v \in L^1(\mu_Q)$ [17, 16]. Under additional conditions on the Hamiltonian system, which hold for the most significant application class in molecular dynamics, one can show that the propagator satisfies the conditions (C1) and (C2) of the preceding section [16].

Figure 1 illustrates the basic idea of the identification algorithm: metastable sets are identifiable via almost constant levels or "sign structure" of the dominant eigenvectors.

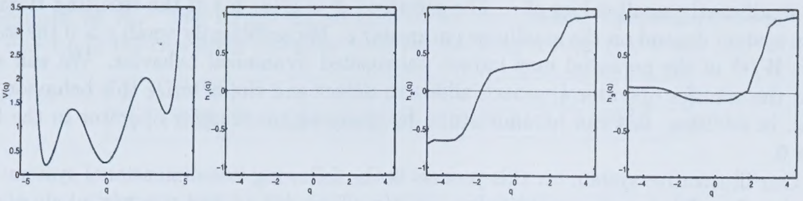


Figure 1: Three well potential and dominant eigenvectors. The choice of β implies the average kinetic energy to be small enough so that the neighborhood around the three potential energy minima are metastable sets of the Markov chain (9). Please observe that the dominant eigenvectors are almost constant on these sets and that the combination of signs of the three eigenvectors—the so–called sign structure—are different on the three sets.

Langevin dynamics. The most popular model for an *open system* with stochastic interaction with its environment is the so-called Langevin model [14]:

$$\dot{q} = p, \qquad \dot{p} = -\nabla_q V(q) - \gamma p + \sigma \dot{W} \tag{10}$$

with some friction constant $\gamma > 0$ and an external force given by a $3N$-dimensional Brownian motion W_t. The stochastic force models the influence of the Brownian motion of the heat bath surrounding the molecular system. The process induced by the Langevin equation (10) leaves the canonical measure μ_{can} corresponding the the inverse temperature β invariant, if the noise and damping constants satisfy $\beta = 2\gamma/\sigma^2$ [14]. The evolution of $u = u(x,t)$ with respect to μ_{can} is governed by the well–known *Fokker–Planck equation* [14]:

$$\partial_t u = \left(\underbrace{\frac{\sigma^2}{2}\Delta_p - p \cdot \nabla_q + \nabla_q V \cdot \nabla_p - \gamma p \cdot \nabla_p}_{=A} \right) u. \tag{11}$$

As a consequence, the operator A is the infinitesimal generator of the semigroup of forward transition operators $\{P_t\}_{t\in\mathbf{R}_0^+}$ acting on $L^1(\mu_{\text{can}})$ with

$$P_t u = \exp(tA)u. \tag{12}$$

Since μ_{can} is invariant, we have $P_t \mathbf{1}_\mathbf{X} = \mathbf{1}_\mathbf{X}$. Moreover, under certain conditions on the potential V, this is the *unique stationary density* and the semigroup $\{P_t\}_{t\in\mathbf{R}_0^+}$ is asymptotically stable [11], i.e., $P_t u \to \mathbf{1}_\mathbf{X}$ for $t \to \infty$ and every density $u \in L^1(\mu_{\text{can}})$. Due to this property, the Langevin equation is the most prominent stochastic model for a heat bath driven relaxation of molecular ensembles to the canonical ensemble.

6 Identifying Dynamical Barriers in MD

In order to illustrate the power of the algorithmic idea presented so far, let us return to the stochastically perturbed Hamiltonian dynamics and consider so-called strong constraining potentials of the form

$$V_\epsilon(q) = U(q) + \frac{1}{\epsilon^2}W(q),$$

for significantly small values of ϵ. The solutions $q^\epsilon = q_t^\epsilon(q_0, p_0)$ of the resulting Hamiltonian system depend on the smallness parameter ϵ. For sufficiently small $\epsilon > 0$ the strong part W/ϵ^2 of the potential may induce unexpected dynamical behavior. We will show that the transfer operator approach allows to detect and characterize this behavior, and that, in addition, this can be understood by studying the transfer operator in the limit $\epsilon \to 0$.

Our illustrative system for this process is the following two-dimensional system: $q = (x, y) \in \Omega = \mathbf{R}^2$, in polar coordinates $x = r(x, y)\cos\phi(x, y)$ and $y = r(x, y)\sin\phi(x, y)$, with potential

$$V_\epsilon(r, \phi) = U(\cos\phi) + \frac{1}{\epsilon^2}W(r, \phi), \qquad W(r, \phi) = \frac{1}{2}\,\omega^2(\phi)\,(r - \pi)^2, \tag{13}$$

with $\omega = \omega(\phi)$ smooth and strictly bounded away from zero. Figure 2 illustrates this potential for the specific choices $U(c) = (c^2 - 1)^2/2$ and $\omega(\phi) = 1 + \exp(-(\phi - \pi)^2/\sigma^2)$ mod 2π with $\sigma = 0.05$, which we will herein discuss in detail.

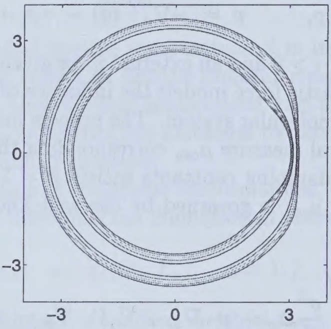


Figure 2: Illustration of the potential V_ϵ for $\epsilon = 0.1$ with U and ω as given in the text above. Contour plot with equipotential curves for $V_\epsilon = 0, 2, 4, \ldots, 10$. Observe that, *on the circle* $r = \pi$, the potential has two saddle points at $\phi = \pi/2, 3\pi/2$, and two global minima, one for $\phi = 0$, the other one in the middle of the narrow pathway at $\phi = \pi$.

In this case the solutions q_t^ϵ display the following feature: the motion is a combination of a slow motion along the circle $r = \pi$ and some significantly faster oscillations in radial direction, normal to the circle, which are induced by the part W/ϵ^2 of the potential. These fast normal oscillations may induce unexpected dynamical behavior. For certain choices of the initial conditions, the amplitude of these normal oscillations is too large to pass the narrow pathway at $\phi = \pi$ (see Fig. 2), *although* the potential V_ϵ has one of its *global minima* at $(r, \phi) = (\pi, \pi)$. One says, that there is a "dynamical barrier" at $\phi = \pi$. Whenever one considers an statistical ensemble of initial states, e.g., the canonical ensemble, the dynamical barrier has the effect that a certain ratio of trajectories will never cross the point $\phi = \pi$ so that it may separate two metastable sets from each other. Whenever, in addition, the average kinetic energy in the ensemble is significantly smaller than the potential energy barrier at the local saddle points at $(r, \phi) = (\pi, \pi/2)$ and $(r, \phi) = (\pi, 3\pi/2)$, these points may also induce separations between metastable sets. Thus, one may expect to find three metastable sets separated by the lines $\phi = \pi/2$, $\phi = \pi$, and $\phi = \pi/2$. We will see below that this eventually is true.

Spectrum of P_ϵ. The largest eigenvalues of the full transfer operator P_ϵ for $\epsilon = 0.1$ and $\tau = 1.5$ in the canonical ensemble with inverse temperature $\beta = 5$ are:

k	1	2	3	4	5	6	...
λ_k	1.0000	0.9938	0.9639	0.9221	0.8441	0.7895	...

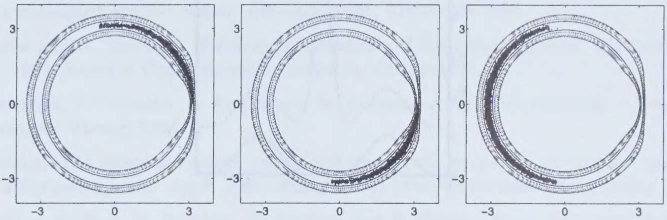


Figure 3: Illustration of the three metastable sets computed via the identification algorithm.

Evaluation of the eigenvectors corresponding to the three dominant eigenvalues yields a decomposition of the inner part of the potential energy well into three metastable sets which are shown in Fig. 3 above. Since it is the common physical belief that metastable sets of molecular motion are separated by saddle points of the potential energy function, our illustrative systems displays an unexpected dynamical behavior.

6.1 Limit dynamics

In order to further understand our observation, we study the motion in the limit $\epsilon \to 0$. Therefore, consider the sequence $q^\epsilon = q_t^\epsilon(q_*^\epsilon, p_*^\epsilon)$ of solutions of (8) with potential V_ϵ given by (13). We also use the alternative notations $q^\epsilon = (x^\epsilon, y^\epsilon) = (r^\epsilon, \phi^\epsilon)$. Let $(q^*, p^*) \in \Omega \times \mathbf{R}^d$ be arbitrary initial values, and $S_\epsilon : \Omega \to \Omega$ the scaling transformation given by $S_\epsilon(x, y) = (R_\epsilon \cos\phi, R_\epsilon \sin\phi)$, with $R_\epsilon = \pi + (r(x, y) - \pi)\,\epsilon$. Consider the family of solutions $q^\epsilon = q_t^\epsilon(S_\epsilon q^*, p^*)$ in a finite time interval $[0, T]$.

In the limit $\epsilon \to 0$, the strong part W/ϵ^2 of the potential will constrain the motion to the minimum of W, i.e., to the manifold $\mathcal{M} = \{q \in \Omega : \quad r(q) = \pi\}$. For every set (q, p) of state variables, we can find the projection q_M of every positions q onto $\mathcal{M}$ via $q_M = \pi\,(\cos\phi(q), \sin\phi(q))$ with normal component $q_N = (r(q) - \pi)\,(\cos\phi(q), \sin\phi(q))$, and decompose the momenta into components tangential and normal to $\mathcal{M}$ at q, i.e.,

$$p = p_M + p_N \quad \text{with} \quad p_N = \frac{p^T q}{q^T q}\, q.$$

According to [15, 2, 1] the limit dynamics q^0 is the strong limit of $q^\epsilon = q_t^\epsilon(S_\epsilon q^*, p^*)$ in $C^1[0, T]$ and is governed by the *limit equation of motion* that—in polar coordinates $q^0 \mapsto (r^0, \phi^0)$—has the form

$$\ddot{\phi}^0 + \frac{1}{\pi^2}\mathrm{grad}_\phi\,(U + \theta\,\omega)\,|_{\phi=\phi^0} = 0, \qquad r^0 = \pi, \tag{14}$$

with initial conditions $\phi^0(0) = \mathrm{arcos}(q_M^*/\pi)$ and $\dot{\phi}^0(0) = (e_\phi(\phi^*)p_M^*)/\pi$, where e_ϕ denotes the unit vector in the direction of ϕ in polar coordinates. The parameter θ

is a constant which is uniquely determined by the initial values via the formula $\theta = \frac{1}{\omega(q^*_M)}\left(\frac{1}{2}\,|p^*_N|^2 + W(q^*)\right)$. Obviously, the limit motion on the circle $\mathcal{M}$ sees an additional potential $\theta\omega$ which represents the influence of the energy that is contained in the motion normal to $\mathcal{M}$, although this motion has vanishing amplitude for $\epsilon \to 0$. For details see [2].

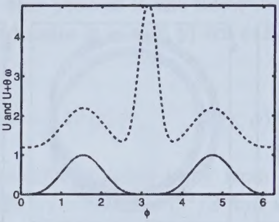


Figure 4: Original potential U on the manifold $\mathcal{M}$ and corrected potential $U + \theta\omega$ for the case shown in Fig. 2 and initial conditions $p^* = (1.5, 1.5)$ and $q^* = (0.1, \sqrt{\pi - 0.1})$.

Figure 4 clearly indicates that the correcting potential $\theta\omega$ induces an additional potential energy barrier at $\phi = \pi$ to the original potential U on the circle $\mathcal{M}$. This perfectly explains the existence of a dynamical barrier at $\phi = \pi$ for small $\epsilon > 0$.

Limit Transfer Operator T_0. It is shown in [20] that the limit transfer operator in the canonical ensemble is given by

$$T_0 u(q) = \int_{\mathbf{R}^d} u\left(q^0_\tau(q_M, p)\right) \mu_P(dp). \tag{15}$$

Obviously, P_0 acts on functions defined on the constraining manifold $\mathcal{M}$.

Spectrum of P_0. The six largest eigenvalues of P_0 for $\tau = 1.5$ with $\beta = 5$ are

k	1	2	3	4	5	6	...
λ_k	1.0000	0.9933	0.9633	0.9246	0.8440	0.7945	...

which is in perfect agreement to the results given above. Evaluation of the eigenvectors corresponding to the three dominant eigenvalues yields a decomposition of the circle $r = \pi$ into three metastable sets ($\phi \in [0, \pi/2] \cup [3\pi/2, 2\pi]$, $\phi \in [\pi, 3\pi/2]$, and $\phi \in [\pi/2, \pi]$). This result perfectly corresponds to the sets computed from the full transfer operator P_ϵ for $\epsilon = 0.1$, see Fig. 3.

References

[1] F. Bornemann. *Homogenization in Time of Singularly Perturbed Mechanical Systems*, volume 1687 of *Lecture Notes in Mathematics, no. 1687.* Springer, Berlin, Heidelberg, New York, 1998.

[2] F. A. Bornemann and C. Schütte. A mathematical approach to smoothed molecular dynamics: Correcting potentials for freezing bond angles. *Physica D*, 102:52–77, 1997.

[3] K. L. Chung. *Lectures from Markov Processes to Brownian Motion*, volume 249 of *A Series of Comprehensive Studies in Mathematics.* Springer, New York, Heidelberg, Berlin, 1982.

[4] M. Dellnitz and O. Junge. On the approximation of complicated dynamical behavior. *SIAM J. Num. Anal.*, 36(2):491–515, 1999.

[5] P. Deuflhard, M. Dellnitz, O. Junge, and C. Schütte. Computation of essential molecular dynamics by subdivision techniques. In P. Deuflhard, J. Hermans, B. Leimkuhler, A. Mark, B. Skeel, and S. Reich, editors, *Computational Molecular Dynamics: Challenges, Methods, Ideas.*, Lecture Notes in Computational Science and Engineering, pages 94–111. Springer Verlag, 1998.

[6] P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte. Identification of almost invariant aggregates in nearly uncoupled Markov chains. *Lin. Alg. Appl.*, 315:39–59, 2000.

[7] M. Freidlin and A. Wentzell. *Random perturbations of dynamical systems.* Springer, New York, London, 1984. Series in Comprehensive Studies in Mathematics.

[8] R. Haberlandt, S. Fritzsche, G. Peinel, and K. Heinzinger. *Molekulardynamik - Grundlagen und Anwendungen.* Vieweg, 1995.

[9] W. Huisinga. The essential spectral radius and asymptotic properties of transfer operators. *submitted to Dynam. Systems Appl.*, 2000. Available via http://www.math.fu-berlin.de/~huisinga.

[10] W. Huisinga, C. Best, R. Roitzsch, C. Schütte, and F. Cordes. From simulation data to conformational ensembles: Structure and dynamics based methods. *To appear in J. Comp. Chem.*, 1998. Available as ZIB Preprint SC-98-36 via http://www.zib.de/bib/pub/pw/.

[11] K. Ichihara and H. Kunita. A classification of the second order degenerate elliptic operators and its probabilistic characterization. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, 30:235–254, 1974.

[12] S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability.* Springer, Berlin, Heidelberg, New York, Tokyo, 1993.

[13] D. Revuz. *Markov Chains.* North–Holland, Amsterdam, Oxford, 1975.

[14] H. Risken. *The Fokker-Planck equation.* Springer, New York, London, 2nd edition, 1996.

[15] H. Rubin and P. Ungar. Motion under a strong constraining force. *Comm. Pure Appl. Math.*, 10:65–87, 1957.

[16] C. Schütte. *Conformational Dynamics: Modelling, Theory, Algorithm, and Application to Biomolecules.* Habilitation Thesis, Fachbereich Mathematik und Informatik, Freie Universität Berlin, 1998. Available as Report SC-99-18 via http://www.zib.de/bib/pub/pw/.

[17] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comput. Phys., Special Issue on Computational Biophysics*, 151:146–168, 1999.

[18] C. Schütte and W. Huisinga. On conformational dynamics induced by Langevin processes. Preprint SC-99-25, Konrad–Zuse–Zentrum, Berlin. Available via http://www.zib.de/MDGroup, 1999.

[19] C. Schütte, W. Huisinga, and P. Deuflhard. Transfer operator approach to conformational dynamics in biomolecular systems. Preprint SC-99-36, Konrad–Zuse–Zentrum, Berlin. Available via http://www.zib.de/MDGroup, 1999.

[20] Ch. Schütte, W. Huisinga, and A. Fischer. On the algorithmic identification of essential degrees of freedom in systems with complicated dynamical behavior. Manuscript in preparation, 2000.

[21] G. Singleton. Asymptotically exact estimates for metatstable Markov semigroups. *Quart. J. Math. Oxford*, 35(2):321–329, 1984.

[22] H. Zhou, S. Wlodek, and J. McCammon. Conformation gating as a mechanism for enzyme specificity. *Proc. Natl. Acad. Sci. USA*, 95:9280–9283, 1998.

Perfect Sampling and Queueing Models*

R.L. Tweedie
Division of Biostatistics
University of Minnesota
Minneapolis, MN 55455
tweedie@biostat.umn.edu

J.N. Corcoran
Dept. of Applied Mathematics
University of Colorado
Boulder, CO 80309
corcoran@colorado.edu

Abstract

We describe an algorithm for simulating "perfect" random samples from the invariant measure of a Harris recurrent Markov chain. The method uses backward coupling of embedded regeneration times, and works most effectively on continuous spaces for stochastically monotone chains, where paths may be sandwiched between "upper" and "lower" processes. An application of the algorithm to bounded storage models is given, and we also develop a random "upper" process for analyzing unbounded storage models.

1 Introduction

There has been considerable recent work on the development and application of algorithms that will enable the "perfect" simulation of the invariant measure π of a Markov chain, sparked by the coupling from the past (CFTP) algorithm in the seminal paper of Propp and Wilson [8]. These ideas have proven effective in many areas such as statistical physics or spatial point processes, and have also been used in operations research on queueing models [4], where they provide simple and powerful alternatives to methods based on iterating transition laws, for example.

The essential idea of the CFTP algorithm, which is given more formally in for example [8, 3, 2], is to find a random epoch $-T$ in the past such that, if we construct sample paths from *every point* in the state space starting at $-T$, then all paths will have coupled successfully by time zero. If at $-T$ we were to draw from π then intuitively, since the paths couple regardless of the state we choose, the resulting common value at time zero also has distribution π.

In Section 2 of this paper we describe a variation on the CFTP algorithm from [2] which can be used when the chain is Harris recurrent, even though in such a model the paths of the chain may not actually meet, due to (for example) the transition laws being continuous. The key idea is to wait until all paths are in some "small set" and by constructing an artificial regeneration for all paths at that time, to ensure that they do couple on an extended state space involving the regeneration time variables. This idea has been used by Murdoch and Green [7] when the whole space is small (so that the chain is uniformly ergodic), and their results can be seen as special cases of our method.

---

*Work supported in part by NSF Grants DMS-9504561 and DMS-9803682, and CRDF Grant RM1-226

These coupling methods are very effective in implementation when the chain has some monotonicity properties, since then one needs only to construct paths from the maximal and minimal elements, using the fact that all other paths are "sandwiched" between these two. When the top and bottom paths are in the small set then one can attempt to regenerate all paths simultaneously. Many chains of interest in operations research satisfy such stochastic monotonicity properties, and in the absence of maximal or minimal elements it is also often the case that we can construct stochastically upper and lower dominating processes to give the desired sandwiching result.

An effective version of the algorithm for monotone chains, which is intuitively appealing, is given in Section 3. In Section 4 we implement the simplified method in practice for a storage system with no attainable minimal element. If the space is bounded then the upper process needed is trivial, although the Harris coupling requires work; when the space is unbounded then the upper process needed is far from trivial, since we need to construct a non-reversible upper process with the correct structure.

2 A Harris chain algorithm

Let us call a set C *small* if the transition law is minorized for some $0 < \varepsilon < 1$ and for some density φ, but only on C: that is,

$$P(x, dy) \geq \varepsilon\varphi(dy), \qquad x \in C. \tag{1}$$

Now consider the Nummelin splitting construction [6], in which we construct sample paths by using the transition law $P(x, \cdot)$ for $x \in C^c$, but each time $x \in C$ we draw a sequence ε_n of i.i.d. $U[0,1]$ random variables, and a sequence V_n of i.i.d. variables with law φ; if $\varepsilon_{n+1} \leq \varepsilon$ then we set $X_{n+1} = V_{n+1}$, and if $\varepsilon_{n+1} > \varepsilon$ then we choose X_{n+1} from the "residual" distribution

$$R(X_n, \cdot) = [P(X_n, \cdot) - \varepsilon\varphi(\cdot)]/[1 - \varepsilon].$$

With this structure we see that that X_{n+1} has the correct marginal distribution [6].

This suggests a method of backward coupling analogous to that in [7]. Let us write the stochastic recursions for $\mathbf{X}$ as

$$X_{n+1} = f(X_n, \xi_n), \;\; n = \ldots, -1, 0, 1, \ldots,$$

where the ξ_n are a doubly infinite sequence of $U[0,1]$ variables: this is described in more detail in [3].

Now draw a doubly infinite set of i.i.d. $U[0,1]$ variables ε_n, independent of the ξ_n. Consider the family of chains $\mathbf{X}^{(x),m}$, with $\mathbf{X}^{(x),m}$ starting from $x \in \mathsf{X}$ at time m. If we can find a time T such that

(i) all of the values $X_r^{(x),-T}$ starting at time $-T$ are simultaneously in C at a time r with $-T < r < 0$, and

(ii) the value of $\varepsilon_r \leq \varepsilon$,

then we can update all of the chains at $r + 1$ with a value drawn from φ; and they will then all have the same value at time zero. Thus T will be a backward coupling time and the common value is a perfect draw from π. The feasibility of finding such a time T which is finite is discussed in [2].

In carrying out this algorithm recursively backward in time, we must as with the Propp-Wilson algorithm take care that we reuse the same uniform random variables at each time; and we must in particular take care that if, at a time point m, we *fail* to regenerate using φ because $\varepsilon_m > \varepsilon$, then every path that enters C at time m from an earlier starting time (before $-T$) also fails to regenerate at m. In other words, the chain we must use is the *bivariate chain*

$$Z_n = (X_n, \varepsilon_n), \qquad n \geq 0 \tag{2}$$

if we are to use the coupler above. This is in direct contrast to the construction in the forward direction, where we need only generate values of ε_n at times when the chain is actually in C.

We note that in practice the implementation will be of value only if we can ensure that the family $\mathbf{X}^{(x),-T}$ all enter the small set C simultaneously (and in a reasonable time). Trivially, this will happen for the uniformly ergodic case if we take $C = \mathsf{X}$: this will lead to the Murdoch-Green algorithm. We see below that this also occurs for monotone chains under some conditions. however, verifying it in non-monotone situations may be computationally impossible except for special cases.

3 Bounding the family $\mathbf{X}^{(x)}$

In order to actually implement this algorithm, we will consider a stochastically monotone chain $\mathbf{X} = \{X_n\}$ on an ordered state space X. By this we mean that there is an order on X such that

$$x \leq y \Longrightarrow f(x, \xi) \leq f(y, \xi)$$

for any realization of ξ, where f is the stochastic recursion generating $\mathbf{X}$.

When the chain is stochastically monotone with an upper and lower bounding process, as described below, we only need to ensure that these processes are in C in order to be able to regenerate using φ. This is analogous to ensuring that the upper and lower processes in the Propp-Wilson algorithm reach a single common point: at that time the "sandwiching" property of the stochastic monotonicity ensures that chains from all starting points have this same common value.

An upper and lower process pair U_n, L_n is defined formally as a pair of processes generated by sequences

$$\begin{array}{llcl} U_0, & U_{n+1} & = & f_U(U_n, \xi_n) \\ L_0, & L_{n+1} & = & f_L(L_n, \xi_n) \end{array}$$

in such a way that two properties hold: the sandwiching property, that for any $x_l \leq x \leq x_u$, we have

$$f_L(x_l, \xi) \leq f(x, \xi) \leq f_U(x_u, \xi);$$

and a "funneling property", that ensures that upper processes starting earlier lie beneath those starting later, and lower processes starting earlier are above those starting later; and lower processes are always below the corresponding upper processes: see [5].

If the space has maximal and minimal points $K, 0$, then U_n, L_n can be taken as the chains starting from these points: that is, we can take $f_L = f_U = f$, with $U_0 = K, L_0 = 0$. When the upper or lower process is not simply the chain itself run from a maximal or minimal point then other processes may be designed, as we now discuss.

We will assume that the small set C is at the "bottom" of the space $\mathsf{X} = [0, K]$, since in this case we do not need to be concerned about the lower process.

Using the splitting technique, let us write $f_R(x,\xi)$ for a function such that the distribution of $f_R(x,\xi)$ is $R(x,\cdot)$, and similarly $f_\varphi(\xi)$ as a function which generates variables with distribution φ. Since we have monotonicity of R, $f_R(x,\cdot)$ is also monotone in x and we will exploit this.

Now let us consider two independent sequences of $U[0,1]$ variables, ε_n and ξ_n, independent for $n = \ldots,-1,0,1,\ldots$; and conduct a coin toss H_n where $H_n = 0$ if $\varepsilon_n \leq \varepsilon$, with $H_n = 1$ otherwise. The stochastic recursive sequence formulation of the chain using splitting [6] is then given by

$$X_{n+1} = \begin{cases} f(X_n,\xi_n) & X_n > c, \\ f_R(X_n,\xi_n) & X_n \leq c, \qquad H_n = 1 \\ f_\varphi(\xi_n) & X_n \leq c, \qquad H_n = 0 \end{cases} \tag{3}$$

We now construct an upper process for the bivariate chain $Z_n = \{X_n, \varepsilon_n\}$ by taking

$$U_{n+1} = \begin{cases} \max\{f(U_n,\xi_n), f_R(c,\xi_n)\} & U_n > c, \qquad H_n = 1 \\ \max\{f(U_n,\xi_n), f_\varphi(\xi_n)\} & U_n > c, \qquad H_n = 0 \\ f_R(U_n,\xi_n) & U_n \leq c, \qquad H_n = 1 \\ f_\varphi(\xi_n) & U_n \leq c, \qquad H_n = 0 \end{cases} \tag{4}$$

It is then obvious that for ordered starting points

$$x \geq y \Longrightarrow \{U_{n+1} | U_n = x\} \geq \{U_{n+1} | U_n = y\} \tag{5}$$

so that funneling holds if $U_0 = K$, the maximal element in the space. It is also clear from (3) that for all x and all sequences ξ_n, ε_n, if $X_0 = x$ and $U_0 = K$ then

$$X_n \leq U_n, \qquad n = 0,1,\ldots \tag{6}$$

and so U_n is indeed an upper process. This leads to:

Monotone Coupling Algorithm:

1. For each $n = 1,2,\ldots$, generate independent $U[0,1]$ variables $\xi_{-n}, \varepsilon_{-n}$;
2. For each $n = 1,2,\ldots$, choose $U_{-n} = K$ and simulate the upper process from this point.
3. Continue until reaching

$$T = \min\{n \geq 1 : U_m^{(K),-n} \in C, \varepsilon_{-n+m} \leq \varepsilon \text{ for some } 1 \leq m \leq n\} \leq \infty; \tag{7}$$

   and let $m(T)$ denote the smallest value of m in (7);

4. For $s = -m(T), -m(T)+1, \ldots, 0$ draw $X_{m(T)+s}^{(x_0),-m(T)}$ according to (3); so that in particular $X_{m(T)+1}^{(x_0),-m(T)}$ has the distribution φ.

This is shown in [2] to be a successful coupling under minimal conditions. although it is rather cumbersome.

In order to simplify matters, suppose we have a special case of the monotone situation above, where the bivariate nature of the chain becomes less relevant, and the algorithm is considerably simplified and more intuitive.

Suppose again that $C = [0, c]$, but now that we can also take $\varphi = P(c, \cdot)$: that is, that we have

$$P(x, \cdot) \geq \varepsilon P(c, \cdot), \qquad x \in C. \tag{8}$$

The constraint (8) is a different relationship than stochastic monotonicity, but is consistent with such monotonicity, and occurs in chains with, for example, appropriate asymptotically random walk behaviour, such as the storage model in Section 4.

Now note that also

$$R(c, \cdot) = [1 - \varepsilon]^{-1}[P(c, \cdot) - \varepsilon\varphi(\cdot)] = P(c, \cdot);$$

and so, regardless of the value of ε_n, we have that when $U_n = x \in C^c$, by monotonicity the draw of U_{n+1} is carried out according to $\max(f(x, \xi), f(c, \xi)) = f(x, \xi)$: that is, rather than constructing a new upper process U_n we just run the chain X_n itself until it reaches the set C, and only then is the value of ε_n relevant, exactly as in the forward splitting construction. Thus in this case the upper process is generated by

$$U^*_{n+1} = \begin{cases} f(U^*_n, \xi_n) & U^*_n > c, \\ f_R(U^*_n, \xi_n) & U^*_n \leq c, \qquad H_n = 1 \\ f_\varphi(\xi_n) & U^*_n \leq c, \qquad H_n = 0 \end{cases} \tag{9}$$

This situation is much easier to implement and we will see it applies to the cases we will consider.

4 Finite Storage Models

We first illustrate our results for a finite storage system on $[0, K]$ with independent and identically distributed exponential replenishments with mean $1/\mu$ at the arrival times of a Poisson process with rate λ. Excessive input above $K < \infty$ is considered overflow and cannot be saved for future use. Between arrivals, content is released deterministically at rate $r(u)$. The Markov chain embedded just prior to arrival times satisfies the PASTA property [1] which ensures that its stationary distribution is identical to the stationary distribution of the continuous time chain.

We shall consider specifically the case $r(u) = \beta u$ for $\beta > 0$. With this release rule, it is easy to see that $\pi(\{0\}) = 0$, so that normal regeneration at zero does not apply.

For $x \in (0, K]$, it can be shown (Lund, private communication) that the density $\pi(x)$ of the stationary distribution π is given by

$$\pi(x) = \frac{x^{(\lambda\beta^{-1}-1)}e^{-\mu x}}{\int_0^K x^{(\lambda\beta^{-1}-1)}e^{-\mu x}dx} \quad \text{for } x \in (0, K] \tag{10}$$

and in general the denominator cannot be integrated.

To execute the algorithm described in Section 3, we need only minorize the process jumps (replenishments) since the release is deterministic once the interarrival times are fixed. For the small set $C = [0, c]$, where $0 \leq c \leq K$, one possible minorization is given by

$$\mathsf{P}(\text{jump to } dh) \geq e^{-\mu c} \cdot e^{\mu c}\mu e^{-\mu h}\mathbb{1}_{\{h \geq c\}}dh \tag{11}$$

where $\mathbb{1}_{\{.\}}$ is the indicator function. Thus we can choose $\varepsilon = e^{-\mu c}$ and $\varphi(dh) = e^{\mu c}\mu e^{-\mu h}\mathbb{1}_{\{h\geq c\}}dh$. Note that in this case we satisfy (8), and the simplified algorithm works.

For this choice of ε and φ, the residual jump probability density $Q(x, x+h)$ is given by

$$Q(x, x+h) = \frac{1}{1-\varepsilon} \begin{cases} \mu e^{-\mu h} & \text{for } h < c - x \\ \mu e^{-\mu h}(1 - e^{-\mu x}) \quad , & \text{for } h \geq c - x \end{cases}$$

where x is the current position of the chain. The residual transition law R is then given by jumping with Q and following the deterministic release path until the time of the next jump.

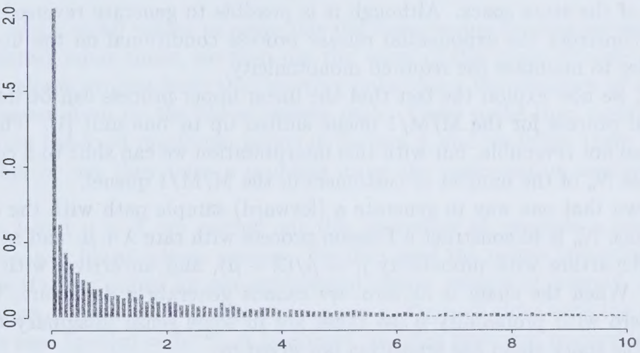
We take $c = 1$ for the simulations. The tractability of the truncated Gamma density given in (10) for the parameter choices $\lambda = 1$, $\mu = 2$, $\beta = 1$ allows us to verify that the algorithm is indeed returning samples from π: in this case

$$\pi(x) = \frac{2e^{-2x}}{1 - e^{-2K}} \quad \text{for } x \in (0, K]. \tag{12}$$

where K is lowered to 1.5.

In Figure 1 we use perfect simulation to show the nature of the density for a parameter combination for which (10) cannot be solved in closed form, and where the mean input is relatively large in relation to other parameters.

Figure 1: 10,000 draws from π for a heavy input dam: $\lambda = 0.4, \mu = 0.1, K = 10$



5 Infinite models and upper processes

At first sight the assumption of the finiteness of the storage system may appear to be critical, since at every negative time-point we begin a sample path generated from the topmost point. In this section, we show that we can adapt the idea (due to Kendall [5]) of an upper bounding random process in order to remove this restriction. David Wilson (personal communication) has also developed a system for doing this.

The general model structure remains as in Section 4, except we now run this model on all of $[0, \infty)$, although due to the release rule it does not reach 0.

To construct an upper bounding process we again consider the storage model with identical input parameters but with the simple unit linear release rate $r(u) \equiv 1$. If this process and our original process are both run forward from the same starting point, with the same driving sequence, it is clear that sample paths of the original exponential release process will stay below the bounding linear process as long as both sample paths are above 1. Thus, set the bottom of the bounding process at 1; that is, whenever the sample path hits 1 (rather than 0) we continue at 1 until the next input time. The original process is allowed to go below 1, as before, and so the linear process on $[1, \infty)$ is always an upper process.

It is known [1] that for $\lambda < \mu$, the stationary distribution π_U for this upper process is

$$\pi_U(1) = \frac{-(\lambda - \mu)}{\mu}, \qquad \pi_U(x) = \frac{-\lambda(\lambda - \mu)}{\mu} e^{(\lambda - \mu)(x-1)}, \qquad x > 1. \tag{13}$$

The method described by Kendall is to start the process at time 0 with this law, then to generate a backward path from this starting (or ending!) point using the time reversal of the upper process in order to get an upper envelope for paths of the original exponential release process.

This is relatively straightforward when the upper process is reversible in the sense that $\pi_U(x)P_U(x,y) = \pi_U(y)P_U(y,x)\ \forall x, y \in \mathsf{X}$, where P_U is the transition law of the bounding process. For then if one chooses the value at time zero using the stationary distribution π_U, one can generate a backward path with the same transition law as the forward process Unfortunately, for the linear process this reversibility breaks down at the bottom of the state space. Although it is possible to generate reversed paths, it is complex to construct the exponential release process conditional on the upper process paths in order to maintain the required monotonicity.

However, we now exploit the fact that the linear upper process can be interpreted as the workload process for the M/M/1 queue shifted up by one unit [1]. This workload process is also not reversible, but with this interpretation we can shift to a consideration of the process N_n of the number of customers in the M/M/1 queuel.

It is known that one way to generate a (forward) sample path with the distribution of the sequence N_n is to construct a Poisson process with rate $\lambda + \mu$: and at each point, generate a departure with probability $p = \mu/(\lambda + \mu)$, and an arrival with probability $q = 1 - p$. When the chain is at zero, we cannot generate a departure, but we just remain at zero with probability p (so these are in some sense imaginary transitions). This countable space chain has transition law given by

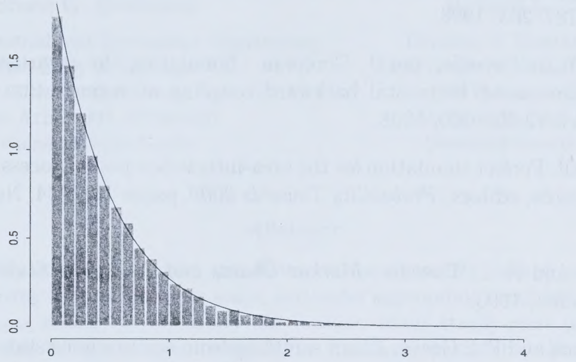
$$P^*(i, i+1) = q, \;\; i \geq 0; \;\; P^*(i, i-1) = p, \;\; i \geq 1; \;\; P^*(0,0) = p$$

with stationary distribution given by $\pi^*(j) \propto [q/p]^j$. It is easily checked that this chain, with "imaginary" transitions, is reversible.

Using this we construct a backward sample path of N_n as follows. Assume there is an arrival point at time 0+, and generate n_0 distributed according to π^*. Now generate the Poisson process backwards; at each successive backward point in the Poisson process we move one step up or down with probability q or p respectively (or at zero, moving up or staying at zero with these probabilities), using the reversibility of the transition laws.

Given this path we then generate a sample path of the workload process conditional on the customer process. This needs care, since at the times of the arrivals the input is precisely the exponential variable that gives the next forward time of departure. Thus we have that within any "busy cycle" we have the information for the whole workload process within that cycle. In the incomplete busy cycle that overlaps with time zero the last n_0 values can be chosen independently of all other variables using the forgetfulness property of the various exponentials involved.

Figure 2: 10,000 draws from π for the infinite storage model: $\lambda = 1$, $\mu = 2$



We now add 1 to all values to make this the linear upper process we want. At each of the (negative) input times, we then use the value of the upper process and run the exponential release process from that starting point using the sequence of arrival times and input values given by the linear process path. These paths then replace the paths generated from the fixed maximum points in Section 4, and we follow them forward and verify whether or not they form a coupling using the minorization method in the set $[0, 1]$.

For $\lambda = 1$ and $\mu = 2$, we carried out this procedure to generate a reverse sample path for the upper linear release process. We then used the points immediately prior to depletions as starting points for our exponential release process of interest. A histogram of 10,000 values drawn from π in this way is given in Figure 2 along with the true density, given in this case by $\pi(x) = 2e^{-2x}$. A backward coupling time was reached in an average of 8.7 embedded time steps with a maximum of 193; this was an extreme with the next largest being 97.

This construction is clearly capable of handling a wide variety of storage models with exponential input and inter-arrival times, and with arbitrary release rules, provided these release rules are faster than linear above some fixed level.

Thus we have an example of a chain where the Doeblin coupling cannot be used directly, but where the use of a small set at the bottom of the state space does lead to perfect sampling from the stationary distribution.

Acknowledgements

We are grateful to Serguei Foss and Wilfrid Kendall for many valuable insights and discussions.

References

[1] S. Asmussen. *Applied Probability and Queues.* John Wiley & Sons, New York, 1987.

[2] J.N. Corcoran and R.L. Tweedie. Perfect simulation for Harris recurrent Markov chains *Ann. App. Prob.*, (to appear).

[3] S.G. Foss and R.L. Tweedie. Perfect simulation and backward coupling. *Stochastic Models*, 14:187–203, 1998.

[4] S.G. Foss, R.L. Tweedie, and J. Corcoran. Simulating the invariant measures of Markov chains using horizontal backward coupling at regeneration times. *Prob. Eng. Inf. Sci*, 12:303–320, 1998.

[5] W.S. Kendall. Perfect simulation for the area-interaction point process. In L. Accardi and C.C. Heyde, editors, *Probability Towards 2000*, pages 218–234, New York, 1998. Springer.

[6] S. P. Meyn and R. L. Tweedie. *Markov Chains and Stochastic Stability.* Springer-Verlag, London, 1993.

[7] D.J. Murdoch and P.J. Green. Exact sampling from a continuous state space. *Scand. J. Statist.*, 1998. to appear (Preprint at http://www.stats.bris.ac.uk/MCMC/).

[8] J.G. Propp and D.B. Wilson. Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms*, 9:223–252, 1996.

Perfect Simulation for Markov Chains Arising From Discrete-Event Simulation

Shane G. Henderson
Dept. of Industrial and Operations Engineering
University of Michigan
1205 Beal Avenue
Ann Arbor, MI 48109-2117
shaneioe@umich.edu

Richard L. Tweedie
Division of Biostatistics, SPH
A460 Mayo Building, Mail Code 303
420 Delaware Street SE
Minneapolis, MN 55455-0378
tweedie@biostat.umn.edu

Abstract

Virtually any discrete-event simulation can be rigorously defined as a Markov chain evolving on a general state space, and under appropriate conditions, the chain has a unique stationary probability distribution. Many steady-state performance measures can be expressed in terms of the stationary probability distribution of the chain.

We would like to apply "coupling from the past" algorithms to obtain samples from the stationary probability distribution of such chains. Unfortunately, the structural properties of the chains arising from discrete-event simulations preclude the immediate application of current coupling from the past algorithms. We describe why this is the case, and extend a class of coupling from the past algorithms so that they may be applied in this setting.

1 Introduction

Virtually any discrete-event simulation can be formally described by a discrete-time Markov chain evolving on a general state space. Intuitively speaking, the chain records a snapshot of the simulation at the time at which events occur. Enough information is included in the state space to ensure that the resulting process is Markov. We will call such chains DES (discrete-event simulation) chains.

The problem of steady-state discrete-event simulation may be loosely defined as that of determining information related to the "long-run" behaviour of the system under study. There is an intimate connection between this problem and the steady-state probability distribution (if it exists) of the discrete-time Markov chain.

One cannot compute the steady-state distribution of a DES chain explicitly except in very special cases. Simulation is a natural alternative. So we are naturally led to the problem of generating samples from the stationary distribution of a DES chain. But how can one obtain such samples if the stationary distribution cannot be computed?

Recently, two methods for generating samples from the stationary distribution of a discrete-time Markov chain have emerged. Fill's algorithm [2] is, at its core, a form of rejection sampler. An alternative to Fill's algorithm is the class of methods collectively known as "coupling from the past" algorithms that were first introduced in [10, 11]. In this paper we will explore the possibility of applying coupling from the past methods to obtain samples from the stationary distribution of a DES chain.

All of these methods operate under the same basic principle which can be described intuitively as follows. Suppose that a stationary version of the Markov chain has been running since time $-\infty$. Then at time 0, its distribution is the stationary distribution. If we can determine the state of this chain at time 0 in finite time, then that state is a sample from the stationary distribution. But how can this state be determined in finite time?

The idea is to generate sample paths of the chain from all possible starting states over the interval $[-T, 0]$, for some time $-T < 0$. If the values of all of these sample paths coincide at time 0 (coalesce), then no matter what state the chain started out in at time $-T$, it will be in the common state at time 0. Therefore, this is a sample from the stationary distribution. If the sample paths have not coalesced by time 0, then one steps back further in time, and generates sample paths from all possible starting states up to time 0, taking care to reuse the previously generated partial sample paths. This process is repeated until all paths have coalesced by time 0.

This form of coupling from the past is known as a vertical coupling. It can be shown [3] that a successful vertical coupling exists if and only if the Markov chain is uniformly ergodic. The algorithm we describe is therefore limited to this context. Other methods have been derived for dealing with chains that are not uniformly ergodic; see [13] for a survey and references.

Coupling from the past can be implemented in many ways, with the choice of implementation depending on the problem context. The method that we explore has strong connections to the theory of regenerative processes as applied in the discrete-event setting [4, 5]. The method is based on so-called minorizations of the transition kernel of the chain, and is an extension of the partitioned multigamma coupler introduced in [9].

Due to space limitations, in this paper we must content ourselves with an overview of the main ideas. Complete details may be found in [6].

This paper is organized as follows. In Section 2 we review the multigamma and partitioned multigamma couplers developed in [9]. Then, in Section 3 we explain via a simple example why these couplers do not apply immediately to DES chains. We then give an intuitive discussion of the coupler that we have developed that we call the "multistep coupler" in Section 4. Next, in Section 5 we describe some simple test examples to show that the multistep coupler performs as we might hope. We will also describe the results from, but not the details for, the example introduced in Section 3. We conclude the paper with a discussion of some interesting research questions, and some comments about how these ideas might prove useful in general discrete-event simulation.

2 Multigamma Coupling

Let $X = (X_n : n \geq 0)$ be a ϕ-irreducible aperiodic Markov chain on state space S, and suppose that the transition kernel P of X satisfies the condition

$$P(x, \cdot) \geq \lambda \varphi(\cdot) \tag{1}$$

for all $x \in S$, some probability distribution φ, and some $\lambda > 0$. Then X is a uniformly ergodic Markov chain (Theorem 16.2.3 of [8]), and has a unique stationary probability distribution π say. We wish to obtain samples from π. One approach to doing so uses the multigamma coupler [9].

Define the *residual kernel*

$$R(x, \cdot) = \frac{P(x, \cdot) - \lambda\varphi(\cdot)}{1 - \lambda},$$

if $\lambda < 1$, and (arbitrarily) $R(x, \cdot) = \delta_x$, the distribution that concentrates its mass at x if $\lambda = 1$. Then

$$P(x, \cdot) = \lambda\varphi(\cdot) + (1 - \lambda)R(x, \cdot). \tag{2}$$

The decomposition (2) suggests that one can generate X_1 conditional on $X_0 = x$ using the composition method (p. 448, [7]) by first generating a Bernoulli random variable Z_0 with $P(Z_0 = 1) = \lambda$. If $Z_0 = 1$, then X_1 is generated from φ, but otherwise it is generated from $R(x, \cdot)$.

The essence of the multigamma coupler is the observation that one may use the same Bernoulli random variable Z_n to generate X_{n+1} from X_n, irrespective of the value of X_n. If $Z_n = 1$, then X_{n+1} can be generated from φ, *independent of the value of* X_n. So if $Z_n = 1$, then all sample paths of X, irrespective of their value X_n at time n, couple at time $n + 1$.

A formal statement of the algorithm is provided in Figure 1 below. Murdoch and Green [9] provide a statement of the multigamma coupling algorithm that is structurally quite different from the one presented in Figure 1. Their algorithm is based on a stochastic recursion, while this one is perhaps closer to the one that might be implemented. Some notes on the algorithm follow.

```
1. Set $T = \lfloor \log U / \log(1-\lambda) \rfloor + 1$, where $U \sim U(0,1)$
2. Generate $X$ with distribution $\varphi$
3. for $n = -T + 1$ to $-1$
4.     Generate $X$ from $R(X, \cdot)$
5. next $n$
6. Return $X$ as a sample from $\pi$
```

Figure 1: Pseudocode for the multigamma coupler.

1. In line 1, we generate the first time $-T$ (before time 0) that $Z_{-T} = 1$. Since $(Z_n : -\infty < n < \infty)$ is an i.i.d. sequence of Bernoulli random variables, T has a geometric distribution, with $P(T = k) = (1 - \lambda)^{k-1}\lambda$, for $k \geq 1$. The random variable U is uniformly distributed on $(0, 1)$, and independent of all other quantities.
2. In line 2, we generate X_{-T+1}. Since $Z_{-T} = 1$, we generate X_{-T+1} from φ.
3. In lines 3 through 5, we successively generate $X_{-T+2}, \ldots, X_0$. Since $Z_n = 0$ for $-T < n \leq -1$, it follows that we must generate from the residual kernels.

Suppose that we resolve to apply this algorithm to the Markov chain $X = (X_n : n \geq 0)$ evolving on state space $S = \{1, 2, 3\}$ with transition matrix

$$P = \frac{1}{4}\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 2 & 2 & 0 \end{bmatrix}.$$

This chain is clearly irreducible and aperiodic, and so has a stationary probability distribution π. Furthermore, the stationary distribution π is easily calculated, so that coupling from the past is not needed. Nevertheless, it is instructive to consider such an approach.

Suppose that we select φ to be the uniform distribution on S. Unfortunately, this chain does not satisfy the condition (1) with this choice of φ. The problem is that there is a 0 in the third row of the transition matrix P, so that we would have to take $\lambda = 0$. However, from state 3, the 2-step transition probabilities to states $1, 2, 3$ are $3/8, 3/8$ and $2/8$ respectively. So while the condition (1) is not satisfied for this chain when φ is the uniform distribution, a relaxed version of the condition is satisfied. In particular, the chain satisfies the following *minorization* condition.

Suppose that one can partition the state space S into a finite collection of sets $A_1, \ldots, A_d$, such that for each $i = 1, \ldots, d$, we have that

$$P^{m(i)}(x, \cdot) \geq \lambda_i \varphi_i(\cdot) \text{ for all } x \in A_i, \tag{3}$$

where $m(i) \geq 1$, $\lambda_i > 0$ and $\varphi_i(\cdot)$ is a probability measure on S. We call equations of the form (3) $m(i)$-minorizations. Note that (1) is a 1-minorization.

Our particular example satisfies the minorization condition by taking $d = 2$, $A_1 = \{1, 2\}$, $A_2 = \{3\}$, $m(1) = 1$, $m(2) = 2$, $\lambda_1 = \lambda_2 = 3/4$ and φ_i to be the uniform distribution for $i = 1, 2$. Many other choices are possible. For example, if we had taken φ to be uniform on the set $\{1, 2\}$ then the stronger condition (1) would have been satisfied.

The minorization assumption, together with irreducibility and aperiodicity assumptions, is sufficient to ensure that X is a Doeblin chain and hence uniformly ergodic, as verified in [6]. Such chains satisfy the condition (1) with the one step transition kernel P replaced by the m-step transition kernel P^m for some $m \geq 1$. The stationary distribution of P^m coincides with that of P, and so one could, in principle, apply the algorithm in Figure 1 to the m-step skeleton chain to obtain samples from π. There are two main reasons why this may not be recommended.

First, the m mentioned earlier may be very large, and so it may be extremely difficult to obtain expressions for the m-step transition kernel P^m. Such expressions seem to be important in implementing the multigamma coupler. Second, it may be faster [9] to develop a coupling algorithm based directly on (3) than on an m-step version of (1). This point is also discussed in [3].

So it is natural to attempt to develop a coupling from the past algorithm based on (3). The "partitioned multigamma coupler" [9] is based on a special case of (3) with $m(i) = 1$ for all i, and so we would like to apply this method in our context.

3 The Need for Multistep Coupling

Unfortunately, it has recently been shown [5] that DES chains will typically not possess a finite number of 1-minorizations that partition the state space, so that the partitioned multigamma coupler cannot be directly applied to most DES chains.

However, in [5] it is also shown that under very reasonable conditions, the chain will typically possess an m-minorization, with $m > 1$. Rather than go into the details of this general result, it is perhaps more helpful to study an example which reveals the main issues. This is the approach that we adopt in the remainder of this section. Our primary goal is to motivate the ideas to follow, by showing that we cannot expect a DES chain to possess the required finite number of 1-minorizations that partition the state space.

Consider a single-server queue with maximum capacity $Q < \infty$ where $Q \geq 2$. Customers arrive to the system according to a renewal process with interarrival time distribution function F_1. If an arriving customer finds Q customers in the system, then the customer does not enter the system. Service times are i.i.d. and independent of the arrival process, with service time distribution function F_0.

Let $X(t)$ denote the number of customers present in the system at time t, and let $T_0(t)$ ($T_1(t)$) denote the time remaining until the next service (arrival, whether the arriving customer will encounter a full system or not). If $X(t) = 0$, we arbitrarily set $T_0(t) = 0$. Let τ_n denote the time of the nth event (arrival or service) for $n \geq 1$, with $\tau_0 = 0$. Let $X = (X_n : 0 \leq n < \infty)$ denote a Markov chain with $X_n = (X(\tau_n), T_0(\tau_n), T_1(\tau_n))$, the state of the system immediately after the nth event.

Recall that we wish to develop minorizations of the transition kernel associated with the chain X.

First consider the set of states $A_1 \stackrel{\Delta}{=} \{x = (q, t_0, t_1) : q = 0, t_0 = 0, t_1 > 0\}$ corresponding to those states where the system is currently empty. The next event must necessarily be an arrival to the system, at which time both the interarrival and service time clocks are reset. So for any $x \in A_1$ and $y = (p, s_0, s_1)$, we have that

$$P(x, dy) = I(p = 1)F_0(ds_0)F_1(ds_1). \tag{4}$$

Thus, the chain X has a 1-minorization on A_1, with $\lambda_1 = 1$, and $\varphi(dy)$ given by the right-hand side of (4).

Now suppose that $x \in A = \{(q, t_0, t_1) : q = 2, 0 < t_0 < t_1\}$, the set of states where there are 2 customers in the system and the next event will be a service completion. If $x \in A$ and $y = (p, s_0, s_1)$, then

$$P(x, dy) = I(p = q - 1)F_0(ds_0)I(s_1 = t_1 - t_0).$$

Notice that the new clock reading s_1 is given by $t_1 - t_0$, the difference in clock readings at the previous step. Therefore, if φ is the probability distribution associated with any 1-minorization of P starting from $x \in A$, it immediately follows that φ must be concentrated on the set $t_1 - t_0 = c$ for some $c > 0$.

Hence, to partition the state space with a finite number of minorizations, we will at a minimum require that F_0 and F_1 are lattice distributions that are concentrated on some set $\{n\delta : 1 \leq n \leq n^*\}$ for some $\delta > 0$ and $n^* < \infty$, so that the set of possible differences in clock readings is finite. If this does not hold, then we cannot hope to partition the state space with a finite number of sets on which 1-minorizations hold.

However, under certain less restrictive conditions on the distributions F_0 and F_1 it *is* possible to partition the state space into a finite number of sets $A_1, \ldots, A_d$ together with m-minorizations where $m > 1$. Indeed, a set of such minorizations is given in [6]. The development there relies on the fact that after 2 transitions, the minorization distributions φ_i need not be concentrated on "knife-edge" sets where the difference in clock readings is fixed.

4 An Intuitive Description of the Multistep Coupler

Suppose that the condition (3) holds for a Markov chain $X = (X_n : n \geq 0)$.

Define the remainder kernels for $i = 1, \ldots, d$ by

$$R_i(x, \cdot) = \frac{P^{m(i)}(x, \cdot) - \lambda_i \varphi_i(\cdot)}{1 - \lambda_i},$$

if $\lambda_i < 1$, and (arbitrarily) $R_i(x, \cdot) = \delta_x$, the distribution that concentrates its mass at x otherwise. Then for $x \in A_i$,

$$P^{m(i)}(x, \cdot) = \lambda_i \varphi_i(\cdot) + (1 - \lambda_i) R_i(x, \cdot).$$

Let

$$m = \max_{i=1}^{d} m(i).$$

The multistep coupler is based on the following approach to generating sample paths of X. Suppose that $X_n \in A_i$. We generate a Bernoulli random variable $Z_n(i)$ with $P(Z_n(i) = 1) = \lambda_i$. If $Z_n(i) = 1$, then $X_{n+m(i)}$ is generated according to φ_i, but otherwise is generated according to $R_i(X_n, \cdot)$. We then use the 1-step transition kernel to generate $X_{n+m(i)+1}, \ldots, X_{n+m}$ (if $m(i) < m$). At time $n + m$, this process is repeated, and so the sample path is generated in steps of size m. Note that we do not generate the "intermediate" values $X_{n+1}, \ldots, X_{n+m(i)-1}$. These values are not needed for the multistep coupler.

We can generate $Z_n(i)$ for $i = 1, \ldots, d$ from a single uniform random variable U_n on $(0, 1)$ by setting $Z_n(i) = I(U_n \leq \lambda_i)$ for each i. To ensure independence of the $Z_n(i)$s across n, we require that $(U_n : -\infty < n < \infty)$ is an i.i.d. sequence.

The multistep coupler generates the first time $n = -mT_1$ before time zero that $Z_n(i) = 1$ for all $i = 1, \ldots, d$. Note that T_1 has a geometric distribution with parameter λ, where $\lambda = \min_i \lambda_i > 0$, so that $P(T_1 = j) = \lambda(1 - \lambda)^{j-1}$ for $j \geq 1$. At time $-mT_1 + m$, we can be sure that every set A_i has coalesced, and so from time $-mT_1 + m$ to time 0 we only need to keep track of d paths, rather than (potentially) uncountably many.

The algorithm is then virtually identical to the partitioned multigamma coupler. In particular, we continue to step back in time in jumps of size mT_i, where $T_1, T_2, \ldots$ are i.i.d. After T_j is generated, we follow d sample paths from time $m \sum_{k=1}^{j} T_k$ to time $m \sum_{k=1}^{j-1} T_k$. We have previously generated sample paths from this latter time forward to time 0, and so we can simply follow those paths up to time 0. If all d paths originating at time $m \sum_{k=1}^{j} T_k$ coincide at time 0, then we terminate and report the common value at time 0 as a sample from π. If not, T_{j+1} is generated and the algorithm continues in a recursive fashion.

The full algorithm is given in [6].

5 Some Examples

In this section we provide 3 simple examples for which the stationary distribution is known and provide numerical results for the multistep coupler. These examples serve both as useful illustrations, and as test cases.

The first two examples are based on a Markov chain $X = (X_n : n \geq 0)$ evolving on a discrete state space $S = \{0, 1, \ldots, n\}$. This example was first described in [1]. The 1-step transition probabilities are defined by $P(x, x) = 1/2$, and $P(x, y) = 1/2n$ for $y \neq x$. The transition matrix P is doubly stochastic, and so the stationary distribution for this chain is discrete uniform on S.

This chain possesses a 1-minorization on S, where φ is the discrete uniform distribution on S, and $\lambda = (n + 1)/2n$. However, to demonstrate the features of the multistep algorithm we will consider more complicated minorizations. In both examples we take $n = 9$.

Example 1 Let $A_1 = \{0, 1, 2\}$, $A_2 = \{3, 4, 5\}$, and $A_3 = \{6, 7, 8, 9\}$.

We could take φ_i as the discrete uniform distribution on S, but it is more instructive to consider a more complicated example. Let φ_i denote the discrete uniform distribution on A_i, for $i = 1, 2, 3$. We can then take $\lambda_1 = \lambda_2 = 1/6$, and $\lambda_3 = 2/9$, while still ensuring that for $x \in A_i$, $P(x, y) \geq \lambda_i \varphi_i(\{y\})$. The residual kernels are then defined appropriately. Note here that we are taking $m(i) = 1$ for all i, so that the full multistep coupling algorithm is not explicitly required, and one could use the partitioned multigamma coupler.

We obtained 5000 samples from the stationary distribution using the multistep coupler. A 95% confidence interval for the mean number of transitions required to obtain a single sample from the stationary distribution was 31.6 ± 0.6. A chi-squared test of goodness of fit, based on 10 bins corresponding to each of the 10 states yielded a test statistic of 9.4. This is well within both the 5% and 95% percentiles (which are 3.325 and 16.919 respectively) of the chi-squared distribution with 9 degrees of freedom. Thus, it would appear that the multistep coupler is performing as it should for this problem.

Example 2 Our second example examines the same Markov chain as did Example 1, and uses the same partition sets A_1, A_2, and A_3. We use the same minorizations as in Example 1 for $x \in A_1, A_2$, but for $x \in A_3$, we take $m(3) = 2$, φ_3 to be the discrete uniform distribution on S, and $\lambda_3 = 130/162$. Indeed, it can be easily verified that $P^2(x, x) = 45/162$, and $P^2(x, y) = 13/162$ for $y \neq x$, and so it follows that for $x \in A_3$ and all $y \in S$, $P^2(x, y) \geq \lambda_3 \varphi(\{y\})$. For $x \in A_3$, it follows that $R_3(x, \cdot)$ is a point mass at x. For this example then, $m = 2$, and $\lambda = 1/6$.

We again obtained 5000 samples from the stationary distribution using the multistep coupler. A 95% confidence interval for the mean number of transitions required to obtain a single sample from the stationary distribution was 39.6 ± 0.8. A chi-squared test of goodness of fit, based on 10 bins corresponding to each of the 10 states yielded a test statistic of 5.4. This is well within both the 5% and 95% percentiles of the chi-squared distribution with 9 degrees of freedom. Thus, it would appear that the multistep coupler is again performing as it should for this problem.

Example 3 Our third example is a Markov chain $X = (X_n : 0 \leq n < \infty)$ on state space S, the half-open unit interval $[0, 1)$. The chain is defined through the recurrence

$$X_{n+1} = [X_n + U_n] \bmod 1,$$

where $(U_n : 0 \leq n < \infty)$ is an i.i.d. sequence of uniformly distributed random variables on $(-1/8, 1/8)$. For this chain, we have that for $x \in S$,

$$P(x, dy) = \begin{cases} 4\,dy & \text{for } y \in B(x, 1/8) \\ 0 & \text{otherwise,} \end{cases}$$

where $B(x, 1/8)$ is the set $\{y : y = (x + w) \bmod 1, -1/8 < w < 1/8\}$, the set of points that can be reached from $x \in S$ in one transition. It is straightforward to verify that the stationary distribution π for this chain is uniform on S.

We take $A_1 = [0, 1/8), A_2 = [1/8, 1/4), A_3 = [1/4, 1/2), A_4 = [1/2, 3/4), A_5 = [3/4, 7/8)$ and $A_6 = [7/8, 1)$.

For $x \in A_1$ and $0 \leq y < 1/8$, we have that

$$P(x, dy) = 4\,dy \geq \lambda_1 \varphi_1(dy),$$

where φ_1 is the uniform distribution on $[0, 1/8)$, and $\lambda_1 = 1/2$. Similarly, we may take φ_2 to be uniform on $[1/8, 1/4)$, φ_5 to be uniform on $[3/4, 7/8)$, and φ_6 to be the uniform distribution on $[7/8, 1)$ together with $\lambda_2 = \lambda_5 = \lambda_6 = 1/2$.

For $x \in A_3 \cup A_4$, we have that $P^2(x, \cdot)$ is a triangular distribution with parameters $(x - 1/4, x, x + 1/4)$, i.e., that

$$P^2(x, dy) = (4 - 16|y - x|)I(|y - x| < 1/4)\, dy.$$

It is then straightforward to show that with $m(3) = 2$, we can take φ_3 to be a triangular distribution with parameters $(1/4, 3/8, 1/2)$ and $\lambda_3 = 1/4$. Similarly, with $m(4) = 2$, we may take φ_4 to be a triangular distribution with parameters $(1/2, 5/8, 3/4)$ and $\lambda_4 = 1/4$.

It is easy to generate iterates from the 1-step transition kernel of this chain, as it is to generate from φ_i for $i = 1, \ldots, 6$, and from the residual kernels $R_i(x, \cdot)$ for $i = 1, 2, 5, 6$. However, generating from the residual kernels $R_3(x, \cdot)$ and $R_4(x, \cdot)$ is somewhat nontrivial. We therefore employ an acceptance rejection method for generating from these distributions, using $P^2(x, \cdot)$ as a proposal distribution.

We again obtained 5000 samples from the stationary distribution using the multistep coupler. A 95% confidence interval for the mean number of transitions required to obtain a single sample from the stationary distribution was 101 ± 2. A chi-squared test of goodness of fit, based on 10 bins corresponding to the 10 subintervals $[k/10, (k + 1)/10)$, $k = 0, 1, \ldots, 9$ yielded a test statistic of 7.7. This is well within both the 5% and 95% percentiles of the chi-squared distribution with 9 degrees of freedom. Thus, it would appear that the multistep coupler is again performing as it should for this problem.

We have also conducted experiments on the DES chain described in Section 3. For full details the reader is referred to [6]. The results there can be summarized by stating that for a queue capacity of size Q, a total of $d = 4Q$ minorizations of the form (3) are required to partition the state space. For $Q = 10$, the expected number of transitions required to identify a sample from the stationary distribution is approximately 1000. Hence, obtaining perfect samples in that case is computationally expensive relative to simulating a single transition of the chain.

6 Conclusions and Future Research

The detailed results in [6] suggest that it is certainly possible to obtain a sample from the stationary distribution of the embedded Markov chain in certain cases. The principal difficulty in obtaining such results lies in constructing the required minorizations. This appears to be a highly nontrivial task, and is, at least currently, a barrier to widespread application. Nevertheless, we have demonstrated that in principle, it is possible to obtain samples from the stationary distribution of the embedded chain.

This naturally leads to a strongly related question, namely, can one obtain samples from the stationary distribution (if it exists) of the continuous-time Markov chain that, sampled at event times, gives rise to the discrete-time Markov chain we have analyzed? If so, we would have a method for solving the initial transient problem, at least for a class of models. Such questions are related to Palm theory for point processes; see, for example, [12].

It is also interesting to ask whether one can avoid the uniform ergodicity requirement implicit in the use of vertical couplings. Some general techniques have been developed to combat this difficulty (an overview and references are given in [13]), but whether these techniques can be successfully applied to this setting remains to be seen.

As noted earlier, there is a strong connection between the theory presented here, and the theory of regenerative simulation of discrete-event dynamic systems (see [5]). Both approaches rely, in the general case, on developing minorizations for m-step transition kernels of a DES chain. If one can apply the multistep coupler developed here, then it appears that one can apply regenerative steady-state simulation methodology. The converse is not necessarily true, because the requirements of the regenerative method are weaker than those of the multistep coupler, at least as presented here.

Given the very attractive statistical properties of estimators constructed from regenerative arguments, one might very well ask why we would want perfect samples from the stationary distribution of the embedded chain. This is a reasonable objection. However, if we are able to leverage these results to obtain perfect samples from the stationary distribution $\tilde{\pi}$ say, of the continuous-time Markov chain alluded to above, then there is certainly potential gain.

To see why, suppose that we are trying to estimate $\alpha = \int f(x)\tilde{\pi}(dx)$, where $\tilde{\pi}$ is the stationary distribution of the continuous-time Markov chain $X = (X(t) : t \geq 0)$, and f is a cost function defined on the state space of X. A regenerative estimator of this quantity is based on a single run of the chain X, and takes the form of a ratio. Typically then, it suffers from "ratio estimation" bias. If this ratio estimation bias is severe and cannot be sufficiently mitigated with bias-reduction techniques, then an alternative approach, based on coupling from the past, might be preferred.

In such a method, we would generate a moderate number, n say, of independent samples from $\tilde{\pi}$ using coupling from the past methods, and then use these samples as initial points in n fixed runlength simulations of the continuous-time Markov chain X. If $Y_1, \ldots, Y_n$ are n sample path averages defined via $Y_i = t^{-1} \int_0^t f(X^i(s))\, ds$, where X^i is the ith replication of the sample path of the chain X and t is the simulation runlength, then the Y_is are i.i.d. *unbiased* estimates of α. We can then estimate α by the average of the Y_is, and confidence intervals are easily constructed. The unbiased nature of this estimator, and the ease with which we can obtain confidence intervals, comes at the cost of generating the n initial points using coupling from the past.

These properties of a coupling from the past type estimator are certainly attractive, and suggest that there is potential value in this approach.

References

[1] Andradóttir S., Calvin, J. M., and Glynn, P. W. (1995) Accelerated regeneration for Markov chain simulations. *Probab. Engrg. Inform. Sci.* **9**, 497–523.

[2] Fill, J. A. (1998). An interruptible algorithm for perfect sampling via Markov chains. *Annals of Applied Probability* **8** 131–162.

[3] Foss, S. G., and Tweedie, R. L. (1998). Perfect simulation and backward coupling. *Communications in Statistics Stochastic Models.* **14** 187–203.

[4] Henderson, S. G., and P. W. Glynn (1999). Derandomizing variance estimators. *Operations Research* **47** 907–916.

[5] Henderson, S. G., and P. W. Glynn (2000). Regenerative steady-state simulation of discrete-event systems. *Under revision.*

[6] Henderson, S. G. and R. L. Tweedie. (2000). Discrete-event steady-state simulation and coupling from the past. *In Preparation.*

[7] Law, A. M., and Kelton, W. D. (2000) *Simulation Modeling and Analysis.* 3rd edn. McGraw Hill, Boston.

[8] Meyn, S. P., and Tweedie, R. L. (1993) *Markov Chains and Stochastic Stability.* Springer-Verlag, London.

[9] Murdoch, D. J., and Green, P. J. (1998) Exact sampling from a continuous state space. *Scand. J. Statist.* **25,** 483–502.

[10] Propp, J. G., and Wilson, D. B. (1996). Exact sampling with coupled Markov chains and applications to statistical mechanics. *Random Structures and Algorithms* **9** 223–252.

[11] Propp, J. G., and Wilson, D. B. (1998). How to get a perfectly random sample from a generic Markov chain and generate a random spanning tree of a directed graph. *Journal of Algorithms* **27** 170–217.

[12] Thorisson, H. (2000). *Coupling, Stationarity and Regeneration.* Springer, New York.

[13] Wilson, D. B. (2000). Layered multishift coupling for use in perfect sampling algorithms (with a primer on CFTP). In N. Madras, ed., *Monte Carlo Methods, volume 26 of Fields Institute Communications*, p. 141–176. American Mathematical Society.

Fault Diagnosis and Consistency in Hybrid Systems

S. Hashtrudi Zad
ECE Department, Concordia University
Montreal, QC, Canada H3G 1M8
shz@ece.concordia.ca

R.H. Kwong, W.M. Wonham
ECE Department, University of Toronto
Toronto, ON, Canada M5S 3G4
kwong, wonham@control.utoronto.ca

Abstract

We study fault diagnosis using discrete–event models and examine the cases in which the discrete–event model is an abstraction of a more complex hybrid automaton. Here an important issue is whether the simplified high–level DES model contains enough information about the plant for the purpose of fault diagnosis. Two notions of consistency between the low–level (hybrid) and high–level (DES) models are introduced. A set of sufficient conditions for consistency and a semi–algorithmic method for obtaining suitable high–level DES models for the low–level hybrid systems are presented.

1 Introduction

Fault diagnosis has been the subject of extensive research. In many diagnostic problems, the plant can be modelled as a discrete–event system (DES). For these problems various methodologies have been proposed (See, e.g., [2, 6, 7]). In [2], we introduced a state–based approach for passive on–line diagnosis of finite–state automata. In this framework, changes in the output sequence resulting from failures are used to determine the **condition** of the system, and hence, detect and isolate failures. In [3], we studied the issue of failure diagnosability.

In many practical cases, the finite–state automaton may be a simplified model (i.e., abstraction) of a hybrid automaton, i.e., a system whose dynamics is described by a combination of differential (or difference) equations *and* a state transition graph (See [1], [4] and references therein). Examples of these systems can be found in flexible manufacturing, process and aerospace systems. Typically, the design of a diagnosis system which is based on a simplified discrete–event model is verified through simulation and test under a large (though finite) number of operating conditions of the (hybrid) system. In this paper, we consider the development of a more systematic approach to the above verification problem. We examine the question of whether a high–level DES contains enough information about the low–level hybrid model which leads us to the introduction of the notion of **consistency** between the high–level and low–level models. We provide a set of sufficient conditions for consistency. The above notion of consistency is somewhat restrictive and, in fact, too strong. In order to be able to build suitable high–level (DES) models for a larger class of hybrid automata, we introduce the notion of **weak consistency**. This will result in a semi–algorithmic method for obtaining suitable high–level models. Weakly consistent models are easier to construct than consistent ones. Unfortunately, in general, no guarantee exists for the successful termination of the above design sequence. For a complete account of our results, the reader can refer to [3].

The paper is organized as follows. In Section 2, we introduce the class of hybrid systems that are of interest to us. Then in Section 3, we extend our framework for fault diagnosis in DES models to hybrid systems. Section 4 discusses the issue of failure diagnosability and introduces the notion of consistency. Section 5 presents a set of sufficient conditions for consistency and a semi–algorithmic method for obtaining suitable (weakly consistent) high–level models for the low–level hybrid systems. Section 6 presents the conclusions.

2 Plant Model

We assume that the plant under control, i.e., plant along with continuous controllers and DES supervisors, can be modelled as a **hybrid automaton** $H = (X, \Sigma, \Theta, \Phi, X_0, Y, \lambda)$, with: $X \subseteq Q \times \mathbb{R}^n$ being the reachable part of the state set; Q the finite set of **discrete states** or **modes**; $\mathbb{R}^n$ the set of continuous–variable components of states; Σ the finite set of events; $\Theta \subseteq Q \times \mathrm{Pwr}(\mathbb{R}^n) \times \Sigma \times \{\psi : \mathbb{R}^n \to \mathbb{R}^n\} \times Q$, the set of transitions; $\Phi = \{f_q : \mathbb{R}^n \to \mathbb{R}^n \mid q \in Q\}$, the set of functions describing the dynamics of the continuous–variable component of the state; $X_0 \subseteq Q_0 \times \Xi_0$, the set of initial conditions ($x(0) \in X_0$), where $Q_0 \subseteq Q$ and $\Xi_0 \subseteq \mathbb{R}^n$; Y the finite set of output symbols; and $\lambda : X \to Y$, the output map.

The **state of the system** $x = (q, \xi)$ has two components: the **mode** $q \in Q$ and the **continuous–variable component** $\xi \in \mathbb{R}^n$. At a mode q, ξ evolves according to $\dot{\xi} = f_q(\xi)$. A transition $(q, \Xi, \sigma, \psi, q')$ becomes enabled if at some time t, the system is at the mode q and the continuous–variable component of the state $\xi(t) \in \Xi$. At this point if the event σ occurs, then, upon transition, the mode of the system becomes q' and the continuous–variable component $\psi(\xi(t))$. We call the function ψ the **assignment function** of the transition. It is assumed that the event set can be partitioned according to $\Sigma = \Sigma_f \dot{\cup} \Sigma_e$. Suppose at a time t, a transition $(q, \Xi, \sigma, \psi, q')$ is enabled. If $\sigma \in \Sigma_f$, then σ will occur unless it is preempted by another event. If $\sigma \in \Sigma_e$, then as long as $\xi \in \Xi$, σ may occur, i.e., σ is simply enabled.

Suppose at $t = t_0$, $x(t_0) = (q_0, \xi(t_0))$. The continuous–variable component will evolve on $\xi_0(t)$ which satisfies $\dot{\xi}_0 = f_{q_0}(\xi_0)$ (with $\xi_0(t_0) = \xi(t_0)$, by definition) until at some $t = t_1$, a transition occurs and the system jumps to the mode q_1 and the continuous–variable component becomes $\psi(\xi_0(t_1))$. Then the continuous–variable component will follow $\dot{\xi}_1 = f_{q_1}(\xi_1)$, with $\xi_1(t_1) := \psi(\xi_0(t_1))$, until at some time $t = t_2$ another transition changes the mode to q_2 and the continuous–variable component becomes $\psi(\xi(t_2))$. The evolution of the mode and the continuous–variable components will continue in a similar fashion after $t = t_2$. The state of the system over the interval $[t_i, t_{i+1}]$ ($i \geq 0$) will be:

$$x_i(t) = (q_i, \xi_i(t)). \tag{1}$$

If $t_i = t_{i+1}$, then it means that after the transition at $t = t_i$, another transition has occurred right away; i.e., two transitions have occurred back–to–back. We assume that in the hybrid system H, only a finite number of transitions may occur back–to–back. We call the sequence of functions $x_i(\cdot)$ given by Eq.(1) a **state trajectory** of the system. Any other sequence of functions, defined on a sequence of closed intervals, that satisfies the dynamics of the system will also be called a state trajectory.

We can think of q, ξ and x as functions of time t *and* the index i. At jump points t_i, in particular, the value of i is required for determining $x = (q, \xi)$. In the following, whenever the value of i is of no importance to the discussion, we simply write $x(t)$, $q(t)$ and $\xi(t)$ to refer to the values of the state and its components, with the understanding that these are the values of the state and its components at time t in some interval.

In this work, we assume that at a mode q, the output $\lambda(x)$ is constant; i.e., λ is a function of q only: $\lambda((q, \xi)) = \lambda((q, \xi'))$ for all $(q, \xi), (q, \xi') \in X$. This simply means that output changes are treated as transitions. For example, a change of output from y^1

to y^2 is a result of some transition $(q_1, \Xi_{12}, \sigma_{12}, \psi_{12}, q_2)$ from a mode q_1 to q_2, with σ_{12} denoting the event corresponding to the output change. Typically, $\psi_{12}(\cdot)$ is the identity map: $\psi_{12}(\xi) = \xi$. Ξ_{12} is the boundary between the regions in $\mathbb{R}^n$ where the output is y^1 and y^2. As a result of this assumption, the continuous–variable component of the state, $\xi(t)$, is defined and known after any output change, and given by the assignment function $\psi(\cdot)$ of the corresponding transition. If the hybrid system does not satisfy the above assumption, it can be transformed to a new system having the same trajectories in which the assumption holds [3].

The model H describes the behaviour of the system in both normal and faulty situations. Suppose there are p **failure modes** $F_1, \ldots, F_p$. The event set Σ includes **failure events**. For brevity, we consider single–failure scenarios only, in other words, at most one failure mode may occur at a time. As with fault diagnosis in finite–state automata, our results can be easily generalized to the case of simultaneous occurrence of two (or more) failure modes [3]. Let $\mathcal{K} := \{N, F_1, \ldots, F_p\}$ denote the condition set of the system. It is assumed that the state set X can be partitioned according to the condition of the system: $X = X_N \dot{\cup} X_{F_1} \dot{\cup} \cdots \dot{\cup} X_{F_p}$. Define the **condition map** $\kappa : X \to \mathcal{K}$ such that for every $x \in X$, $\kappa(x)$ is the condition of the system at the state x: $\kappa(x) = N$ if $x \in X_N$, and $\kappa(x) = F_i$ if $x \in X_{F_i}$ $(i \in \{1, \ldots, p\})$. Also extend the definition of κ to the subsets of X: $\kappa(z) = \cup\{\kappa(x) \mid x \in z\}$, for any $z \subseteq X$. The condition of a hybrid system changes as a result of failure (or perhaps repair or recovery) events. Therefore, κ is only a function of q (mode), not ξ (continuous–variable component).

In fault diagnosis for hybrid systems (similar to the case of finite–state automata in [2]), we use the output sequence $(y_1 y_2 y_3 \ldots)$ to find the condition of the system. If a transition (other than output changes) is observable, then by defining new output symbols and introducing new transitions (if necessary), the information about the occurrence of the transition can be transferred and included in the output map [3]. Therefore, from now on, we shall assume that the output sequence contains all of the information available about the system.

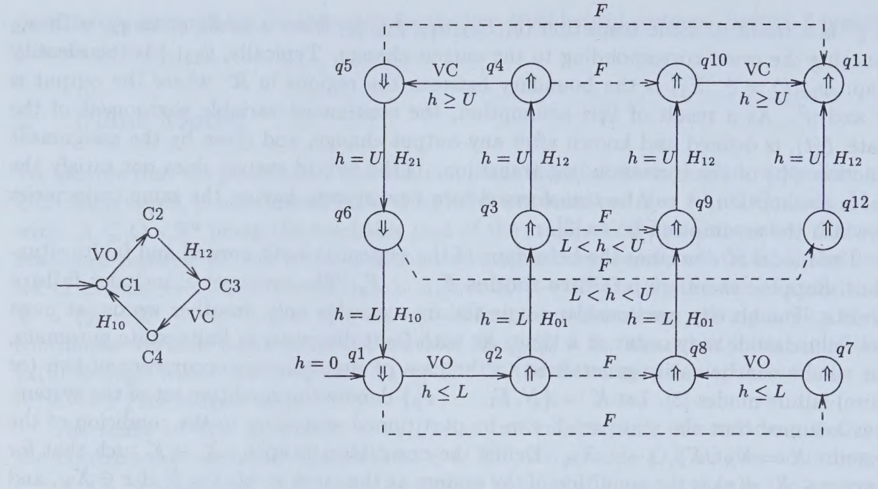
Example - Water Tank: Water is charged into a tank at a (volume) flow rate of c_0 (when the control valve is open). The tank is discharged at a rate of $c = k\sqrt{h}$, where k is a constant and h the water level. When water is at or below the lower threshold L, the controller opens the valve and when water is at or above U, the controller closes the valve. Initially, the tank is empty and the valve is closed. Let A be the cross–sectional area of the tank and $h_m := (c_0/k)^2$. When the valve is open

$$\dot{h} = \frac{c_0 - c}{A} = \frac{k}{A}(\sqrt{h_m} - \sqrt{h}), \tag{2}$$

and when it is closed

$$\dot{h} = -\frac{c}{A} = -\frac{k}{A}\sqrt{h}. \tag{3}$$

It is assumed that the valve may become stuck–open (the event F). Let us denote the regions $\{0 \leq h \leq L\}$, $\{L < h < U\}$ and $\{U \leq h \leq h_m\}$ by H_0, H_1 and H_2. We assume that a sensor measures water level and generates one of the symbols H_0, H_1 and H_2 accordingly. Let H_{ij} denote the transition from region H_i to H_j. The state transition graph of the controller is depicted in Fig. 1(a). Here VO and VC refer to the commands "valve open" and "valve closed", respectively. We assume that the state of the controller is available for fault diagnosis. Therefore, for this system $Y \subseteq \{H_0, H_1, H_2\} \times \{\mathrm{C1}, \mathrm{C2}, \mathrm{C3}, \mathrm{C4}\}$. A hybrid automaton for the water tank is shown in



(a)

(b)

Figure 1: Example. (a) Controller, (b) Hybrid Automaton.

Fig. 1(b). In this system, $\Sigma_f = \{\text{VO}, \text{VC}\}$, $\Sigma_e = \{F\}$, and the assignment functions of the transitions ($\psi(\cdot)$) are all identity maps. Note that whenever output changes, the system mode changes as well. In Fig. 1(b), ⇑ and ⇓ refer to the equations (2) and (3), respectively. In this system, $\xi = h$ and $Q = \{q1, \ldots, q12\}$. □

3 Diagnoser

In our framework, the diagnoser for the hybrid system H is a system (not necessarily finite–state) that takes the output sequence of the system $(y_1 y_2 \ldots y_k)$ as input and generates at its output an estimate of the condition of the system at the time that y_k was generated. Specifically, based on the output sequence up to y_k, a set $z_k \in 2^{Q \times \mathbb{R}^n} - \{\emptyset\}$ is calculated to which x must belong at the time that y_k was generated. $\kappa(z_k)$ will be the estimate of the system's condition. Upon observing y_{k+1}, z_k will be updated to z_{k+1}.

Definition 1 For any two states $x, x' \in Q \times \mathbb{R}^n$, we say x' is **output–adjacent** to x and write $x \Rightarrow x'$ if $\lambda(x) \neq \lambda(x')$ and there exists a trajectory of finite duration, satisfying the hybrid system's dynamics, starting at x and ending at x' such that for all $x'' \neq x'$ belonging to the trajectory: $\lambda(x'') = \lambda(x)$. □

We define the diagnoser to be a machine (not necessarily finite–state) $D = (Z \cup \{\underline{z}_0\}, Y, \zeta, \underline{z}_0, \hat{\mathcal{K}}, \kappa)$, where $Z \cup \{\underline{z}_0\}$, Y and $\hat{\mathcal{K}} \subseteq 2^{\mathcal{K}} - \{\emptyset\}$ are the state, event and output sets of D. $\underline{z}_0 := (z_0, 0)$ is the initial state, with $z_0 \in 2^X - \{\emptyset\}$. $Z \subseteq 2^X - \{\emptyset\}$ and is not necessarily finite. $\zeta : Z \cup \{\underline{z}_0\} \times Y \to Z$ is the transition (partial) function and $\kappa : Z \cup \{\underline{z}_0\} \to \hat{\mathcal{K}}$ the output map, with the definition of the condition map, κ, extended according to: $\kappa(\underline{z}_0) := \kappa(z_0)$. Each diagnoser state z, except $\underline{z}_0$, is identified with a nonempty subset of X. $\underline{z}_0$, however, is considered to be different from other states of the diagnoser because the update law ζ at $\underline{z}_0$ is different from that at other states. To

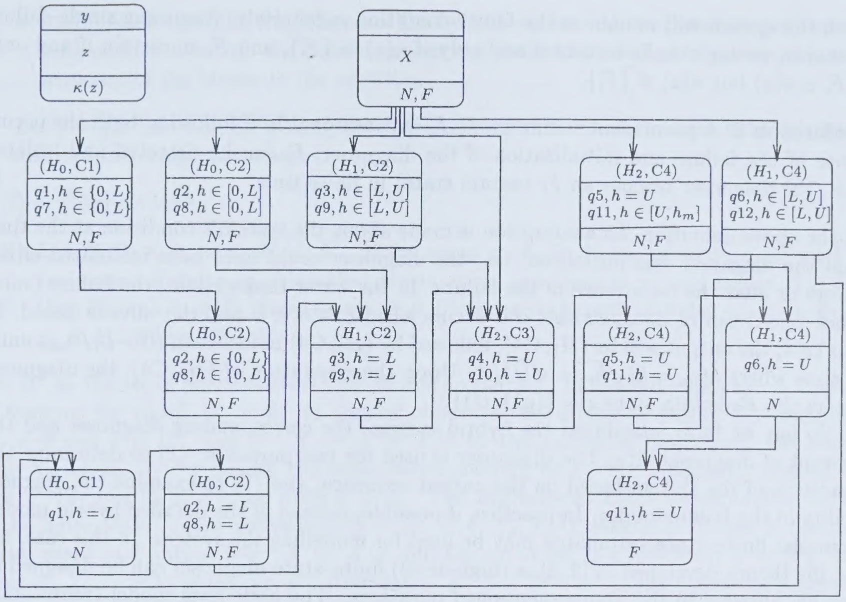

y
z
κ(z)
X
N, F
(H₀, C1)
q1, h ∈ {0, L}
q7, h ∈ {0, L}
N, F
(H₀, C2)
q2, h ∈ [0, L]
q8, h ∈ [0, L]
N, F
(H₁, C2)
q3, h ∈ [L, U]
q9, h ∈ [L, U]
N, F
(H₂, C4)
q5, h = U
q11, h ∈ [U, h_m]
N, F
(H₁, C4)
q6, h ∈ [L, U]
q12, h ∈ [L, U]
N, F
(H₀, C2)
q2, h ∈ {0, L}
q8, h ∈ {0, L}
N, F
(H₁, C2)
q3, h = L
q9, h = L
N, F
(H₂, C3)
q4, h = U
q10, h = U
N, F
(H₂, C4)
q5, h = U
q11, h = U
N, F
(H₁, C4)
q6, h = U
N
(H₀, C1)
q1, h = L
N
(H₀, C2)
q2, h = L
q8, h = L
N, F
(H₂, C4)
q11, h = U
F

Figure 2: Example. Diagnoser.

distinguish $\underline{z}_0$ from other states, it has been identified with a pair $(z_0, 0)$ (rather than a subset). The diagnoser state transition $z_{k+1} = \zeta(z_k, y_{k+1})$ is given by:

$$\begin{aligned} z_1 &= z_0 \cap \lambda^{-1}(\{y_1\}) \\ z_{k+1} &= \{x \mid \lambda(x) = y_{k+1} \ \& \ (\exists\, x' \in z_k : \ x' \Rightarrow x)\}, \quad k \geq 1. \end{aligned}$$

Therefore z_{k+1} will be the set of states, having output y_{k+1}, that are reachable from the states in z_k using trajectories along which the output is y_k. The initial state estimate, z_0, contains the information available about the state of the system at the time that the diagnoser is initialized, before the reading of sensors begins. Similar to the case of finite–state automata, usually $z_0 = X$.

Example - Water Tank (Cont'd): The diagnoser for the water tank system is given in Fig. 2 (assuming $z_0 = X$). In this example, the diagnoser has a finite number of states. □

4 Diagnosability and Consistency

In [3], a failure mode F_i of a discrete–event system is considered **diagnosable** if there exists an integer $N_i > 0$ such that the failure can be detected and isolated after the occurrence of at most N_i events following both the occurrence of the failure and initialization of the diagnoser. We define diagnosability in hybrid systems similarly except that we replace "bounded number of events" with "finite time". Let us assume that the failure modes of the hybrid system are **permanent** which means that if a failure occurs, then the system will remain in the faulty condition indefinitely. Assuming single–failure scenario, we say z is F_i–certain if and only if $\kappa(z) = \{F_i\}$, and F_i–uncertain if and only if $F_i \in \kappa(z)$ but $\kappa(z) \neq \{F_i\}$.

Definition 2 A permanent failure mode F_i is **diagnosable** if following both the occurrence of the failure and initialization of the diagnoser, F_i can be detected and isolated (i.e., the diagnoser reaches an F_i–certain state) in finite time. □

In the above definition, no assumption is made about the system's condition at the time that the diagnoser was initialized; i.e., the diagnoser could have been initialized either before or after the occurrence of the failure. In the water tank system, the failure (valve stuck–open) can be detected only if it occurs when $L < h < U$ and the valve is closed. In this case, the output will be $(H_1, \mathrm{C4})$, followed by $(H_2, \mathrm{C4})$ in less than $(U-L)/\dot{h}_{\min}$ units of time where $\dot{h}_{\min} = k(\sqrt{h_m} - \sqrt{U})/A$. Upon the generation of $(H_2, \mathrm{C4})$, the diagnoser enters the F–certain state $z = \{(q11, U)\}$.

So far, we have introduced the hybrid system, the corresponding diagnoser and the concept of diagnosability. The diagnoser is used for two purposes: (1) to determine the condition of the system based on the output sequence; and (2) to examine the diagnosability of the failure modes. In practice, if possible, instead of the detailed hybrid model, a simpler finite–state automaton may be used for modelling the system. In this case, using the theory developed in [2, 3], a (high–level) finite–state diagnoser can be designed to provide answers to the above–mentioned questions. The high–level model (finite–state automaton) is simpler than the low–level hybrid model and contains less information about the system. At this point, we are not concerned with how the high–level model is obtained.

Definition 3 The low–level and high–level models of the system as described above are said to be **consistent** (for the purpose of fault diagnosis) if and only if (i) the low–level and high–level diagnosers are equivalent; i.e., they generate the same condition estimate in response to a given output sequence of the system, and (ii) a failure mode is diagnosable with respect to the low–level model if and only if it is diagnosable with respect to the high–level model. □

Therefore, we say the models are consistent if the diagnosers provide identical answers to questions (1) and (2) above.

5 Consistent Model Reduction

In this section, we introduce a particular set of partitions of the state set X of the hybrid system H. If this set is nonempty, then we show how a consistent high–level finite–state DES model for the hybrid system can be constructed using any of these partitions.

Let $\pi = \{B^1, \ldots, B^p\}$ be a finite partition of X, with B^i denoting its blocks. Also let $P : X \rightarrow X/\pi = \overline{X}$ be the canonical projection that maps every state x to its coset in X/π. In this work, for simplicity, we do not distinguish between a coset $\overline{x} \in \overline{X}$ and the corresponding block $P^{-1}\overline{x}$. Suppose $x(\cdot)$ is a trajectory consisting of a sequence of functions $x_i(\cdot)$. As the state of the system evolves on the trajectory, it crosses or passes through an ordered set of blocks of π which we call **the projection of the trajectory** (on X/π). We assume that the projection of every trajectory is a *countable* ordered set, i.e., a sequence of blocks (In general, this may not be the case [3]).

- Assumption (a): The trajectories of the hybrid system and the partitions, including $\ker\lambda$ and $\ker\kappa$, satisfy the condition that the projection of any trajectory $x(\cdot)$, is a sequence of the blocks of the partition.

Let $\pi = \{B^1, B^2, \ldots, B^{|\pi|}\}$ be a *finite* partition of the state set X of the hybrid system H with the following properties:

(I) $\pi \leq \ker\kappa \wedge \ker\lambda$;

(II) Let $B, B' \in \pi$, with $B \neq B'$. If for some $x_1 \in B$, there exists $x_1' \in B'$ and a trajectory of finite duration that starts at x_1 and ends at x_1' and lies in $B \cup B'$, then for any $x_2 \in B$, there exists $x_2' \in B'$ and a trajectory of finite duration from x_2 to x_2' lying in $B \cup B'$.

Let Π^* be the set of partitions with the above properties. Timed automata are examples of systems for which $\Pi^* \neq \emptyset$. In general, however, Π^* might be empty [3]. Suppose $\Pi^* \neq \emptyset$ and $\pi \in \Pi^*$. Let $P : X \to \overline{X} = X/\pi$ be the canonical projection. For every $x \in X$, let $[x] := Px$ denote the coset corresponding to x. For simplicity, instead of $x \in P^{-1}\overline{x}_1$ (for $\overline{x}_1 \in \overline{X}$), we write $x \in \overline{x}_1$. The **high–level DES** corresponding to H is the finite–state Moore automaton $\overline{G} = (\overline{X}, \overline{\Sigma}, \overline{\delta}, \overline{X}_0, Y, \overline{\lambda})$, where $\overline{X}$, $\overline{\Sigma}$ and Y are the state, event and output sets. $\overline{X} := X/\pi$ and the event set is a singleton $\overline{\Sigma} = \{\overline{\sigma}\}$, for some $\overline{\sigma}$. $\overline{X}_0 := PX_0$ is the set of initial states. $\overline{\delta} : \overline{X} \times \overline{\Sigma} \to 2^{\overline{X}}$ is the transition function and $\overline{\lambda} : \overline{X} \to Y$ the output map. They are defined as follows. For all $\overline{x}_1, \overline{x}_2 \in \overline{X}$, by definition $\overline{x}_2 \in \overline{\delta}(\overline{x}_1, \overline{\sigma})$ if and only if there exist x_1 and x_2 with $x_1 \in \overline{x}_1$ and $x_2 \in \overline{x}_2$ and a trajectory of finite duration from x_1 to x_2 lying in $\overline{x}_1 \cup \overline{x}_2$. For all $\overline{x} \in \overline{X}$, $\overline{\lambda}(\overline{x}) := \lambda(x)$ for any $x \in \overline{x}$. Similarly, we define $\overline{\kappa} : \overline{X} \to \mathcal{K}$ according to $\overline{\kappa}(\overline{x}) = \kappa(x)$ for any $x \in \overline{x}$. Since $\pi \leq \ker\lambda \wedge \ker\kappa$, $\overline{\lambda}$ and $\overline{\kappa}$ are well–defined. It was assumed in Sec. 2 that the output sequence contains all of the information available about the system through observation. This information is retained at the high level by the output map $\overline{\lambda}$. Therefore we do not need to differentiate between events at the high level (or the low level for that matter) and hence, let the event set of the high–level model be a singleton. It can be shown that if x is output–adjacent to x' ($x, x' \in X$), then $[x]$ is output–adjacent [1] to $[x']$, and conversely, if $\overline{x}$ is output–adjacent to $\overline{x}'$ ($\overline{x}, \overline{x}' \in \overline{X}$), then every state in $\overline{x}$ is output–adjacent to some state in $\overline{x}'$. We define the canonical projection of a subset $z \subseteq X$ to be $Pz := \bigcup\{[x] \mid x \in z\}$. Let $\overline{D}$ be the (high–level) diagnoser designed for the high–level DES $\overline{G}$ (following the procedure given in [2]), with $\overline{z}_0 := Pz_0$ as its initial state estimate (z_0 is the initial state estimate in the (low–level) diagnoser for the hybrid system).

Definition 4 The low–level and high–level diagnosers are **equivalent** if for any output sequence, they produce identical sequences of estimates for the system's condition. □

Theorem 1 The low–level diagnoser D and the high–level diagnoser $\overline{D}$ are equivalent. □

Next we show the equivalence of high–level and low–level diagnosabilities. First we make two assumptions the validity of which depends on the partition π and the dynamics of the hybrid system. Later, we will examine the assumptions in more detail.

---

[1] In the Moore automaton $\overline{G}$, we say $\overline{x}$ is output–adjacent to $\overline{x}'$ and write $\overline{x}' \Rightarrow \overline{x}$ if $\overline{\lambda}(\overline{x}) \neq \overline{\lambda}(\overline{x}')$ and either $\overline{\delta}(\overline{x}', \overline{\sigma}) = \overline{x}$ or there exist $l \geq 1$ and $\overline{x}_1, \ldots, \overline{x}_l$ with $\overline{\lambda}(\overline{x}_i) = \overline{\lambda}(\overline{x}')$ $(1 \leq i \leq l)$ such that $\overline{\delta}(\overline{x}', \overline{\sigma}) = \overline{x}_1$, $\overline{\delta}(\overline{x}_i, \overline{\sigma}) = \overline{x}_{i+1}$ $(1 \leq i \leq l-1)$, and $\overline{\delta}(\overline{x}_l, \overline{\sigma}) = \overline{x}$ [2].

- Assumption (A): The projection of any low–level trajectory of finite duration on the high–level state space is a finite sequence of states.
- Assumption (B): Suppose that the projection of a low–level trajectory on the high–level state space is a single state, say $\overline{x}$, with $\overline{\delta}(\overline{x}, \overline{\sigma}) \neq \emptyset$. Then the trajectory must be of finite duration.

Assumption (A) states that over a finite interval of time, only a finite number of events can occur at the high–level DES. Note that assumption (A) is stronger than assumption (a), previously introduced in this section.

Theorem 2 A failure mode is diagnosable with respect to the low–level (hybrid) model if and only if it is diagnosable with respect to the high–level (DES) model. □

Remark 1 Loosely speaking, assumptions (A) and (B) mean: "finite number of events is equivalent to finite duration". This should not come as a surprise because at the low level a failure is diagnosable if it can be detected and isolated in a "finite time" while at the high level, failure detection and isolation is expected to happen in a "finite number of events". If the two definitions are to become equivalent, then we can expect some sort of equivalence between the "real time" and "logical time". □

Remark 2 Consider a state $\overline{x}$ of a finite–state automaton for which $\overline{\delta}(\overline{x}, \overline{\sigma}) \neq \emptyset$ for all $\overline{\sigma} \in \overline{\Sigma}_s \subseteq \overline{\Sigma}$, with $\overline{\Sigma}_s \neq \emptyset$. In fault diagnosis for finite–state automata, it is implicitly assumed that if the state of the automaton reaches $\overline{x}$, then at some point in the future one of the events in $\overline{\Sigma}_s$ will occur (and therefore the information about this transition and the possible future transitions can be used for diagnosis). This is exactly what assumption (B) implies. □

Example - Water Tank (Cont'd): The reader can verify that the partition $\pi = \{B_1, \ldots, B_{16}\}$, with $B_1 = \{x \mid q = q1, h = 0\}$, $B_2 = \{x \mid q = q2, h \in [0, L]\}$, $B_3 = \{x \mid q = q3, h \in [L, U)\}$, $B_4 = \{x \mid q = q3, h = U\}$, $B_5 = \{x \mid q = q4, h = U\}$, $B_6 = \{x \mid q = q5, h = U\}$, $B_7 = \{x \mid q = q6, h \in (L, U]\}$, $B_8 = \{x \mid q = q6, h = L\}$, $B_9 = \{x \mid q = q1, h = L\}$, $B_{10} = \{x \mid q = q7, h = 0\}$, $B_{11} = \{x \mid q = q7, h = L\}$, $B_{12} = \{x \mid q = q8, h \in [0, L]\}$, $B_{13} = \{x \mid q = q9, h \in [L, U]\}$, $B_{14} = \{x \mid q = q10, h = U\}$, $B_{15} = \{x \mid q = q11, h = [U, h_m]\}$, $B_{16} = \{x \mid q = q12, h \in [L, U]\}$, satisfies assumptions (I) and (II), and hence, belongs to Π^*. The corresponding high–level model and diagnoser are contructed in [3]. The low–level and high–level diagnosers produce identical estimates for the system's condition in response to the system's output sequence. Note that the trajectories of the hybrid system satisfy assumptions (A) and (B). □

In general, it is very difficult to verify property (II) for a given partition (or even verify whether $\Pi^* \neq \emptyset$). This makes the construction of high–level models difficult. In the following, we introduce another method for building high–level models which is based on partitions that only have to satisfy property (I) and assumption (B). These DES models are significantly easier to construct. We will show that the condition estimates generated by the resulting high–level diagnosers will be more conservative; however, the diagnosers may still be used for fault diagnosis and for verifying diagnosability. Here we are taking an approach similar to that proposed in [5] by Kurshan *et al.* for developing a semi–algorithmic method for verifying properties of digital circuits.

Usually, it is very difficult to calculate X, the reachable part of the state space of a hybrid system. However, determining a superset of X might be easier. Let X^* denote one such superset; hence $X \subseteq X^* \subseteq Q \times \mathbb{R}^n$. Let $\hat{\pi} \leq \ker\lambda \wedge \ker\kappa$ and $P_{\hat{\pi}} : X^* \to X^*/\hat{\pi}$ be the canonical projection. Based on this partition, let us define the finite–state Moore automaton $\hat{G} = (\hat{X}, \hat{\Sigma}, \hat{\delta}, \hat{X}_0, Y, \hat{\lambda})$, with $\hat{X}$, $\hat{\Sigma}$ and Y as the state, event and output sets. $\hat{X} := X^*/\hat{\pi}$ and $\hat{\Sigma}$ is a singleton; i.e., $\hat{\Sigma} = \{\hat{\sigma}\}$, for some $\hat{\sigma}$. Let $\hat{X}_0 := PX_0$ be the set of initial states, $\hat{\delta} : \hat{X} \times \hat{\Sigma} \to 2^{\hat{X}}$ the transition function and $\hat{\lambda} : \hat{X} \to Y$ the output map. For $\hat{x}_1, \hat{x}_2 \in \hat{X}$, let $\hat{x}_2 \in \hat{\delta}(\hat{x}_1, \hat{\sigma})$ if and only if there exist $x_1 \in \hat{x}_1$ and $x_2 \in \hat{x}_2$ and a trajectory of finite duration (satisfying the hybrid system's dynamics) from x_1 to x_2 lying in $\hat{x}_1 \cup \hat{x}_2$. Note that since $\hat{\pi}$ does not necessarily satisfy property (II), the block $\hat{x}_2$ may not be reachable from every state in $\hat{x}_1$. For all $\hat{x} \in \hat{X}$, $\hat{\lambda}(\hat{x}) := \lambda(x)$ for any $x \in \hat{x}$. Similarly define $\hat{\kappa} : \hat{X} \to \mathcal{K}$ according to $\hat{\kappa}(\hat{x}) := \kappa(x)$ for any $x \in \hat{x}$ [2]. Since $\hat{\pi} \leq \ker\lambda \wedge \ker\kappa$, $\hat{\lambda}$ and $\hat{\kappa}$ are well–defined. $\hat{G}$ is a high–level model for H. In this automaton, a transition from a state $\hat{x}_1$ to another state $\hat{x}_2$ indicates the existence of a trajectory from block $\hat{x}_1$ to block $\hat{x}_2$ lying entirely in $\hat{x}_1 \cup \hat{x}_2$. It can be shown that if x is output–adjacent to x' ($x, x' \in X^*$), then $P_{\hat{\pi}}x$ is output–adjacent to $P_{\hat{\pi}}x'$ [3].

Define the canonical projection of a subset $z \subseteq X$ to be $P_{\hat{\pi}}z := \cup\{P_{\hat{\pi}}x \mid x \in z\}$. Let $\hat{D}$ be the high–level diagnoser designed based on $\hat{G}$, with $\hat{z}_0 := P_{\hat{\pi}}z_0$ as its initial state estimate. Let $\hat{z}_k$ denote the state of $\hat{D}$ following the generation of the output symbol y_k.

Theorem 3 Let z_k and $\hat{z}_k$ denote the states of the low–level diagnoser D and high–level diagnoser $\hat{D}$, respectively. Then $\kappa(z_k) \subseteq \hat{\kappa}(\hat{z}_k)$. □

The theorem states that $\hat{D}$ produces more conservative estimates for the system's condition than D. Hence, $\hat{D}$ is less accurate than D. If $\hat{D}$ enters an F_i–certain state, so will D (but not vice versa).

Let us suppose assumption (B) holds for $\hat{\pi}$, namely, if the projection of a low–level trajectory in X on the high–level state space $X/\hat{\pi}$ is a single state $\hat{x}$, with $\hat{\delta}(\hat{x}, \hat{\sigma}) \neq \emptyset$, then the trajectory must be of finite duration. Since calculating X is difficult, we may verify assumption (B) for the projection of the trajectories in X^* on $X^*/\hat{\pi} = \hat{X}$ which is sufficient to guarantee assumption (B) (for the trajectories in X).

Theorem 4 If a failure mode is diagnosable with respect to the high–level model $\hat{G}$, then it must be diagnosable with respect to the low–level hybrid model H. □

The above theorem implies that if all failure modes are diagnosable with respect to the high–level model, then the low–level model can be replaced with the high–level one for the purpose of fault diagnosis. However, if a failure turns out to be undiagnosable at the high level, it may still be diagnosable at the low level. In this case, the analysis based on the high–level model is inconclusive. At this point, we may replace the existing partition with another (perhaps finer) partition, develop a new high–level model, and repeat the analysis. We should also consider the possibility that the failure may be undiagnosable at the low level. In these cases, information from the high–level model (e.g., fault–indeterminate cycles [3]) may assist us in proving the undiagnosability of the failure mode at the low level by providing clues about trajectories that may lead the low–level diagnoser into fault–uncertain states.

Based on the above discussion, we introduce the notion of *weak consistency.*

---

[2] λ and κ were earlier defined on X. Here we have assumed that their domain is even larger and includes X^*. This is typically the case. The reachable part of the state set is usually a subset of the domains of the output and condition maps.

Definition 5 The low–level and high–level models of the system are said to be **weakly consistent** (for the purpose of fault diagnosis) if and only if [i] the estimates for the system's condition generated by the low–level diagnoser are subsets of the estimates given by the high–level diagnoser ($\kappa(z_k) \subseteq \hat{\kappa}(\hat{z}_k)$), and [ii] if a failure mode is diagnosable with respect to the high–level model, then it must be diagnosable with respect to the low–level model. □

The models G and $\hat{G}$ are weakly consistent. The reader can refer to [3] for an example.

In the high–level model $\hat{G}$, $\hat{x}_2 \in \hat{\delta}(\hat{x}_1, \hat{\sigma})$ for some $\hat{x}_1, \hat{x}_2 \in \hat{X}$ if there exists a trajectory from an $x_1 \in \hat{x}_1$ to some $x_2 \in \hat{x}_2$, lying entirely in $\hat{x}_1 \cup \hat{x}_2$. Verifying this property (i.e., existence of the trajectory) is significantly simpler than property (II) of the partitions in Π^*. In fact, this property should be investigated only on the boundary between $\hat{x}_1$ and $\hat{x}_2$; specifically, we have to see if a trajectory crosses the boundary [3].

6 Conclusion

In this paper, we study the cases in which the DES model used for fault diagnosis is an abstraction of a more complex hybrid automaton. Here an important issue is whether the simplified high–level model (DES) contains enough information about the plant for the purpose of fault diagnosis. We introduce two notions of "consistency" between the high–level (DES) and low–level (hybrid) models. A set of sufficient conditions for consistency and a semi–algorithmic method for obtaining suitable high–level DES models for low–level hybrid systems are presented. As with other problems in formal verification, obtaining a general systematic solution is extremely difficult (if not impossible). However, it may be possible to enlarge the class of systems that can be formally analyzed by developing semi–algorithmic methods, along with useful mathematical and software tools.

References

[1] Automatica, Special Issue on Hybrid Systems, vol. 35, no. 3, 1999.

[2] S. Hashtrudi Zad, R.H. Kwong and W.M. Wonham, "Fault diagnosis in discrete-event systems: Framework and model reduction," *Proc. 37th IEEE Conf. Decision Contr.*, Tampa, FL, USA, Dec. 1998, pp. 3769–3774.

[3] S. Hashtrudi Zad, "Fault diagnosis in discrete–event and hybrid systems," Ph.D. thesis, University of Toronto, Toronto, Canada, Sept. 1999.

[4] IEEE Transactions on Automatic Control, Special Issue on Hybrid Control Systems, vol. 43, no. 4, 1998.

[5] R. P. Kurshan and K. L. McMillan, "Analysis of digital circuits through symbolic reduction," *IEEE Trans. Computer–Aided Design*, vol. 10, no. 11, pp. 1356-1371, 1991.

[6] F. Lin, "Diagnosability of discrete event systems and its applications," *Discrete Event Dynamic Systems*, vol. 4, pp. 197-212, 1994.

[7] M. Sampath, R. Sengupta, S. Lafortune, K. Sinnamohideen and D. Teneketzis, "Diagnosability of discrete-event systems," *IEEE Trans. Automat. Contr.*, vol. 40, no. 9, pp. 1555-1575, 1995.

Invariance Under Scaling of Time Bounds in Timed Discrete-Event Systems

Sean E. Bourdon, W.M. Wonham
Systems Control Group
Dept. of Electrical and Computer Engineering
University of Toronto
Toronto, ON M5S 3G4
bourdon,wonham@control.toronto.edu

M. Lawford
Dept. of Computing and Software
McMaster University
Hamilton, ON L8S 4L8
lawford@mcmaster.ca

Abstract

A family of affine mappings on the time bounds for timed discrete-event systems is introduced. It is shown that unless these affine maps are in fact linear, then the timed activity transition graph (TATG) of an arbitrary timed discrete-event system (TDES) may not be preserved under the scaling operation. Moreover, it is shown that when the scaling is linear, the TATG is always preserved under scaling. We examine some applications of the result, including state space reduction and make connections to suboptimal supervisory controller synthesis and dense time scaling. Finally, we briefly discuss topics for future study including extensions to model-checking and efficient representation of TDES.

1 Introduction

Several paradigms have been proposed in order to introduce greater modeling flexibility and realism to the Ramadge-Wonham framework of discrete-event systems. Among these is the timed discrete-event systems (TDES) framework of Brandin and Wonham [6] in which the passage of time (in the global sense) is marked by a special event, *tick*. In this theory, time is uniformly quantized and the quantum used represents the elapsed time between successive *ticks* of an integer valued clock. The framework is meant to realistically model sampled systems. Dense time systems, while sometimes mathematically satisfying, are necessarily idealized.

When modeling a real-time system in the TDES framework for the purposes of implementing supervisory control, a faster *tick* rate lends itself to increased accuracy. However, in any physical system, there are limitations on the *tick* rate that may be used for the model. These constraints on the *tick* rate generally arise from two sources. The first of these is hardware limitations. For example, regardless of the technology used, our sampling rate is always bounded. The second constraint on the *tick* rate stems from the well-known state explosion problem. Namely, the size of the state space for a TDES model can grow rapidly as the *tick* rate increases. Hence, a TDES with a high *tick* density may not be very amenable to computation.

One way in which a model's suitability can be judged is via its timed activity transition graph (TATG). A given TDES **G** can be represented graphically using a timed transition graph, which displays all possible sequences of events that can occur within the definition of **G**, including the *tick* transitions. In order to obtain the TATG from the TTG, we merely need to suppress all *tick* transitions. Thus, the TATG incorporates the constraints imposed on the event sequences of **G** by their time bounds.

In Section 3 of this paper, we focus on a specific family of transformations of TDES in which the time bounds for the activities in the original TDES **G** are affinely scaled to produce a new TDES $\mathbf{G}'$. We first consider the case where the *tick* rate is increased in the scaled system. It will be shown that when the time bounds are linearly scaled the TATGs for **G** and $\mathbf{G}'$ will be the same. Moreover, we shall prove that when the scaling is affine but not linear, the TATG for $\mathbf{G}'$ may not necessarily coincide with that of **G**. These first results can assist in determining how a model should change with the introduction of faster components or better processors into the system. We then consider the effect of scaling a system so that *ticks* become slower and discuss application to model reduction. It is also shown that scaling commutes with composition of TDES, so that scaling can be used in a compositional model reduction scheme of the type described in [12] and [11].

Next, we add to the set of results from Section 3 and discuss connections to related work. The results obtained here allow us to relate our work to that of [8] on suboptimal supervisory control.

In Section 4, we examine a few interesting properties that arise in the limit as *ticks* become dense in the real line. Specifically, it is shown that the original definition of TDES breaks down, motivating the need for minor semantic changes.

The final section of the paper summarizes the results presented in the two preceding sections and discusses potential future topics of study.

2 Timed Discrete-Event Systems

As in [6], we model a TDES **G** in two steps. The first is to construct an activity transition structure $\mathbf{G_{act}} := (A, \Sigma, \xi, a_o)$, which is simply a finite deterministic automaton [13]. In this definition, A is the set of activities, $a_o \in A$ is the initial activity, and Σ represents the set of system events. The partial function $\xi : A \times \Sigma \rightarrow A$ is called the activity transition function and we write $\xi(a, \sigma)!$ whenever $\xi(a, \sigma)$ is defined.

The second step in constructing a TDES consists of introducing time into the ATG. In a TDES, time is measured via the *tick* of a discrete global clock, relative to which the enablement of each of the events in Σ is defined. Specifically, to each event $\sigma \in \Sigma$ we assign upper and lower time bounds $u_\sigma \in \mathbb{N}_o \cup \{\infty\}$ and $l_\sigma \in \mathbb{N}_o$, respectively, which satisfy $l_\sigma \leq u_\sigma$. Here, $\mathbb{N}_o := \mathbb{N} \cup \{0\}$, where $\mathbb{N}$ denotes the set of positive integers. An event $\sigma \in \Sigma$ must remain continuously enabled for at least l_σ, but no more than u_σ, *ticks* of the clock in order to occur (provided its occurrence is not preempted by the occurrence of another competing event). In the sequel, we refer to $\Sigma_{tim} := \{(\sigma, l_\sigma, u_\sigma) : \sigma \in \Sigma\}$ as the set of timed events of **G**.

Following [6], we give the TDES **G** an automaton structure. For this, we need some further notation. Let $\Sigma_{spe} := \{\sigma \in \Sigma : u_\sigma < \infty\}$ and $\Sigma_{rem} := \{\sigma \in \Sigma : u_\sigma = \infty\}$. We say that an event σ is prospective (remote) if $\sigma \in \Sigma_{spe}$ ($\sigma \in \Sigma_{rem}$). It is clear from the above definitions that $\Sigma = \Sigma_{spe} \dot\cup \Sigma_{rem}$ and that a remote event has no deadline attached to it. Next, we introduce the special event *tick* to denote the passage of one unit of time in our global clock, and write $\Sigma_t := \Sigma \cup \{tick\}$. Finally, we define the *timer interval*, T_σ, for an event σ via

$$T_\sigma := \begin{cases} [0, u_\sigma], & \sigma \in \Sigma_{spe} \\ [0, l_\sigma], & \sigma \in \Sigma_{rem} \end{cases},$$

where $[j,k]$ denotes the set of integers i satisfying $j \le i \le k$.

We are now able to formally write $\mathbf{G} := (Q, \Sigma_t, \delta, q_o)$. The state set in this case is defined to be $Q = A \times \prod_{\sigma \in \Sigma} T_\sigma$. Similarly, we have that the initial state $q_0 := (a_o, \{t_{\sigma 0} : \sigma \in \Sigma\})$, where $t_{\sigma 0} := \max(T_\sigma)$. Now, suppose $q = (a, \{t_\sigma : \sigma \in \Sigma\})$. We have that $\delta(q, \sigma)!$ if and only if

- $\sigma = tick$ and $(\forall \tau \in \Sigma_{spe})\ \xi(a, \tau)!$ implies $t_\tau > 0$; or
- $\sigma \in \Sigma_{spe}$, $\xi(a, \sigma)!$, and $0 \le t_\sigma \le u_\sigma - l_\sigma$; or
- $\sigma \in \Sigma_{rem}$, $\xi(a, \sigma)!$, and $t_\sigma = 0$

When one of the above conditions is satisfied, we have that $q' = \delta(q, \sigma) := (a', \{t'_\sigma : \sigma \in \Sigma\})$ where either

- $\sigma = tick$ implies $a' := a$ and $t'_\tau := \begin{cases} \max(0, t_\tau - 1), & \xi(a, \tau)! \\ t_{\tau 0}, & \text{otherwise} \end{cases}$
- $\sigma \in \Sigma$ implies $a' := \xi(a, \sigma)$ and $t'_\tau := \begin{cases} t_\tau, & \xi(a', \tau)! \text{ and } \tau \ne \sigma \\ t_{\tau 0}, & \tau = \sigma \text{ or } \neg(\xi(a', \tau)!) \end{cases}$

Now, given an event set, or alphabet, Σ_t we define the set Σ_t^* to be the set of all finite strings of elements in Σ_t, including the empty string ε. It is straightforward to extend the definition for the transition function δ so that $\delta : Q \times \Sigma_t^* \to Q$. Specifically, given $q \in Q$, $\sigma \in \Sigma_t$, and $s \in \Sigma_t^*$, we set $\delta(q, s\sigma) := \delta(\delta(q, s), \sigma)$. Of course, the definition of ξ can similarly be extended. A language over Σ_t is simply a subset of Σ_t^*. The language consisting of strings generated by a TDES $\mathbf{G}$ is called the *closed behaviour* of $\mathbf{G}$ and is denoted by $L(\mathbf{G})$. Using the above definitions, we can write $L(\mathbf{G}) = \{s \in \Sigma_t^* : \delta(q_o, s)!\}$. We also define the projection operator $P_t : \Sigma_t^* \to \Sigma^*$ inductively via

$$P_t(\varepsilon) := \varepsilon$$

$$P_t(\sigma) := \begin{cases} \sigma, & \sigma \in \Sigma \\ \varepsilon, & \sigma = tick \end{cases}$$

$$P_t(s\sigma) := P_t(s)P_t(\sigma).$$

Thus, P_t simply erases all occurrences of the *tick* event in any string $s \in \Sigma_t^*$.

Since a TDES $\mathbf{G}$ has an automaton structure, it can be uniquely represented using a labeled digraph. Such a representation will be referred to as a timed transition graph (TTG) in the sequel. Similarly, we call the graphical representation of $\mathbf{G_{act}}$ the activity transition graph (ATG) of $\mathbf{G}$. A third useful representation of $\mathbf{G}$ is called the timed activity transition graph (TATG) and is the graphical representation of the untimed language $P_t(L(\mathbf{G}))$. Throughout the paper, we assume that all automata and graphical representations are deterministic in their transition structures and minimal state. This ensures that there is a one-to-one correspondence between the automaton, language, and graphical characterizations of $\mathbf{G}$. We denote by $\mathbf{G_{tact}}$ the automaton with the property that $L(\mathbf{G_{tact}}) = P_t(L(\mathbf{G}))$. These definitions in hand, we are now ready to introduce the notion of scaling timed discrete-event systems.

3 Invariance of TATGs Under Scaling of TDES

The goal of this section is to introduce a three parameter family of scaling functions on TDES. The scaling functions are affine transformations on the time bounds of the events in Σ. We choose to work with affine rather than linear scaling functions mainly for flexibility. While linear scaling functions allow for expansion and contraction of time intervals, affine scaling functions provide additional capacity for resizing intervals and also permit translations along the real-time axis. Moreover, when applying inverse scaling to TDES for the purposes of aggregation, affine scaling functions allow for rounding of time bounds and hence approximation methods. This in turn makes scaling applicable to a wider class of systems. We begin by showing that only under linear scaling of time bounds can we guarantee *a priori* the invariance of the TATG of an arbitrary TDES $\mathbf{G}$ under scaling.

3.1 Scaling Timed Discrete-Event Systems

Consider the TDES $\mathbf{G} = (Q, \Sigma_t, \delta, q_o)$ with $\Sigma_t = \Sigma \cup \{tick\}$ and whose timed events belong to the set $\Sigma_{tim} = \{(\sigma, l_\sigma, u_\sigma) : \sigma \in \Sigma\}$.

Definition 1 *Suppose a, b_1, and b_2 are integers. Then $\mathbf{G}' =: S_{a,b_1,b_2}(\mathbf{G})$ is an affine scaling of $\mathbf{G}$ if $\mathbf{G}'_{\mathbf{act}} = \mathbf{G}_{\mathbf{act}}$ and $\Sigma'_{tim} = \{(\sigma, l'_\sigma, u'_\sigma) : \sigma \in \Sigma\}$ where*

$$l'_\sigma := al_\sigma + b_1$$

$$u'_\sigma := \begin{cases} au_\sigma + b_2, & \sigma \in \Sigma_{spe} \\ \infty, & \sigma \in \Sigma_{rem} \end{cases}$$

Moreover, the affine scaling is well-defined if the following conditions are satisfied:

1) $a > 0$

2) $0 \le l'_\sigma \le u'_\sigma$, for every $\sigma \in \Sigma$

3) for every $q \in Q$ and every string $s \in \Sigma^* - \{\varepsilon\}$, $\delta(q, s) \ne q$

The first two conditions say that an affine scaling is well-defined if the scaling factor $a > 0$ and (l'_σ, u'_σ) forms an admissible set of time bounds for each event $\sigma \in \Sigma$, given the choice of offsets b_1 and b_2. The third condition is included for technical reasons. It prevents "non-Zeno" behaviour in which an infinite number of transitions are executed in a finite time interval. This condition can only be violated if $\mathbf{G}_{\mathbf{act}}$ contains a loop whose events *all* have lower time bounds that are mapped to 0 in $\mathbf{G}'$. For future reference, we denote the set of all TDES defined over the alphabet Σ by $\mathbb{G}(\Sigma)$, from which we define the subsets $\mathbb{S}_{a,b_1,b_2}$ by

$$\mathbb{S}_{a,b_1,b_2} := \{\mathbf{G} \in \mathbb{G}(\Sigma) : S_{a,b_1,b_2}(\mathbf{G}) \text{ is well defined}\}.$$

Note that in this notation, we have $\mathbb{S}_{a,0,0} = \mathbb{G}(\Sigma)$ for any integer $a > 0$.

Theorem 2 below constitutes the main result of the paper. It states that the only well-defined affine scalings which preserve the TATG of an arbitrary TDES $\mathbf{G}$ are those which are *linear* – namely, those for which the offsets b_1 and b_2 are identically zero. Moreover, the result states that the TATG will remain the same, regardless of the scaling factor used. Thus, from an event-sequence perspective, refinement of the original model provides no additional insight. This consequence is somewhat reminiscent of the famous Shannon Sampling Theorem for continuous time systems.

Theorem 2 *Given integers $a > 0$, b_1 and b_2, let $\mathbf{G}' := S_{a,b_1,b_2}(\mathbf{G})$ for each $\mathbf{G} \in \mathbb{S}_{a,b_1,b_2}$. Then $L(\mathbf{G}'_{\mathbf{tact}}) = L(\mathbf{G}_{\mathbf{tact}})$ for every TDES $\mathbf{G} \in \mathbb{S}_{a,b_1,b_2}$ if and only if $b_1 = b_2 = 0$.*

The proof of the result is much too long to present here. We merely sketch the main ideas employed and refer the interested reader to [5] for details. Necessity of the condition $b_1 = b_2 = 0$ is easily shown using examples. However, proving the sufficiency of this condition is significantly more challenging. Using the definition of an arbitrary TDES $\mathbf{G}$, it can be shown that if $s = s_0\, tick\, s_1\, tick \ldots \in L(\mathbf{G})$ then $s' = s_0\, tick^n\, s_1\, tick^n \ldots \in L(\mathbf{G}')$, where $\mathbf{G}' := S_n(\mathbf{G})$ and $S_n := S_{n,0,0}$. Since $P_t(s) = P_t(s')$, we have that $L(\mathbf{G}_{\mathbf{tact}}) \subseteq L(\mathbf{G}'_{\mathbf{tact}})$. The reverse subset inclusion is similarly proved using the definition of TDES although it is more difficult to deal with *tick* transitions in this direction.

While an analogous result to that of Theorem 2 is obvious for dense time systems such as the timed automata of Alur and Dill [2], it is somewhat counterintuitive in the TDES framework due to the discretization of time. Consider as an example the TDES $\mathbf{G}$ and $\mathbf{G}'$ whose ATG $\mathbf{G}_{\mathbf{act}}$ is shown in Figure 1.

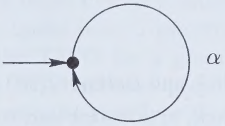


Figure 1: ATG $\mathbf{G}_{\mathbf{act}}$ for the example

Assume for $\mathbf{G}$, we have $\Sigma_{tim} := \{(\alpha, 1, 1)\}$ while $\Sigma'_{tim} := \{(\alpha, 2, 2)\}$. Suppose the *tick* frequency is one *tick* per second for $\mathbf{G}$ while $\mathbf{G}'$ operates with a *tick* rate of two *ticks* per second. The first occurrence of α must take place between the first and second *ticks* in $\mathbf{G}$. Namely, the first occurrence of α must take place in the real-time interval $(1, 2)$. Similarly, we find that the second occurrence of α takes place in the interval $(2, 3)$. Hence, the time between successive α transitions can be arbitrarily small in $\mathbf{G}$. Compare this to the TDES $\mathbf{G}'$ in which the first occurrence of α takes place between the second and third *tick*. Thus, this first occurrence of *tick* takes place in the real-time interval $(1, 1.5)$. It is easy to see that the second occurrence of α must lie in the real time interval $(2, 2.5)$ in $\mathbf{G}'$. So, we find that successive occurrences of α are subject to a relaxation time of at least 0.5 second in $\mathbf{G}'$. This can be seen in Figure 2 below. Note that despite this difference Theorem 2 says that $\mathbf{G}$ and $\mathbf{G}'$ share the same TATG. This phenomenon can lead to difficulties when composing subsystems operating with different granularities of time [5].

Before extending the result in different directions and discussing applications, we present an equivalent characterization of the sufficient condition appearing in Theorem 2. To this end, we make two final definitions before introducing Lemma 3. First, we define the function $tickcount : \Sigma_t^* \to \mathbb{N}_o$ which simply counts the number of occurrences of *tick* in a given string $s \in \Sigma_t^*$. Formally, for any $\sigma \in \Sigma_t$ and $s \in \Sigma_t^*$, we have

$$tickcount(\varepsilon) = 0$$

$$tickcount(\sigma) = \begin{cases} 0, & \sigma \in \Sigma \\ 1, & \sigma = tick \end{cases}$$

$$tickcount(s\sigma) = tickcount(s) + tickcount(\sigma).$$

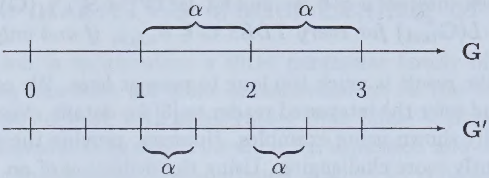


Figure 2: Real time axis

Next, for each $n \geq 2$, we define a new projection operator $P_n : \Sigma_t^* \to \Sigma_t^*$ inductively as follows:

$$P_n(\varepsilon) = \varepsilon$$

$$P_n(\sigma) = \begin{cases} \sigma, & \sigma \neq tick \\ \varepsilon, & \sigma = tick \end{cases}$$

$$P_n(s\sigma) = \begin{cases} P_n(s), & \sigma = tick \text{ and } tickcount(s\sigma) \not\equiv 0 \pmod n \\ P_n(s)\sigma, & \sigma \neq tick, \text{ or } \sigma = tick \text{ and } tickcount(s\sigma) \equiv 0 \pmod n, \end{cases}$$

where in the above $\sigma \in \Sigma_t$ and $s \in \Sigma_t^*$. Thus, given a string $s \in \Sigma_t^*$, the mapping P_n will keep every n^{th} occurrence of *tick* while deleting all other instances. We are now in a position to state our next result.

Lemma 3 *For every* $n \in \mathbb{N}$, $L(\mathbf{G}) = P_n(L(S_n(\mathbf{G})))$.

Notice that the sufficiency of the condition $b_1 = b_2 = 0$ in Theorem 2 immediately follows by application of P_t to both sides of the equality in Lemma 3. That $P_t(L(\mathbf{G})) = P_t(L(S_n(\mathbf{G})))$ implies $L(\mathbf{G}) = P_n(L(S_n(\mathbf{G})))$ can easily be derived using the results found in [5].

3.2 Inverse Scaling and State Space Reduction

The scaling result presented in the last section is useful in integrating faster components into a real-time system. However, inverse scaling, in which the *tick* rate experiences an n-fold *decrease*, has more applicability from a practical point of view. Below, we will show that a TDES can be inversely scaled provided the (finite) time bounds for each activity are divisible by a common integer n. In such a case, the scaled system $\mathbf{G}'$ satisfies the property $L(\mathbf{G}'_{\mathbf{tact}}) = L(\mathbf{G}_{\mathbf{tact}})$. We now make this discussion more precise.

Once again, we suppose the activity set for the TDES $\mathbf{G}$ is Σ and that the corresponding timed events are $\Sigma_{tim} = \{(\sigma, l_\sigma, u_\sigma) : \sigma \in \Sigma\}$. Now consider the mapping S^{-1}_{a,b_1,b_2}. If $\mathbf{G}' := S^{-1}_{a,b_1,b_2}(\mathbf{G})$, then it is clear that $\mathbf{G}'_{\mathbf{act}} = \mathbf{G}_{\mathbf{act}}$ and $\Sigma'_{tim} = \{(\sigma, l'_\sigma, u'_\sigma) : \sigma \in \Sigma\}$, where

$$l'_\sigma := \frac{1}{a}(l_\sigma - b_1)$$

$$u'_\sigma := \begin{cases} \frac{1}{a}(u_\sigma - b_2), & \sigma \in \Sigma_{spe} \\ \infty, & \sigma \in \Sigma_{rem} \end{cases}$$

Hence, the mapping S^{-1}_{a,b_1,b_2} also affinely scales the time bounds of the activities in $\mathbf{G}$. In this case, we will say that $\mathbf{G}'$ is well-defined if $0 \leq l'_\sigma \leq u'_\sigma$, $l'_\sigma \in \mathbb{N}_0$ for every $\sigma \in \Sigma$, and $u'_\sigma \in \mathbb{N}_0$ for each $\sigma \in \Sigma_{spe}$. By analogy with our earlier work, we define the set $\mathbb{S}^{-1}_{a,b_1,b_2}$ to consist of all TDES $\mathbf{G} \in \mathbb{G}(\Sigma)$ such that $S^{-1}_{a,b_1,b_2}(\mathbf{G})$ is well-defined. Using these definitions, we have the following corollary of Theorem 2.

Corollary 4 *Suppose* $a > 0$, b_1 *and* b_2 *are integers. Then* $L(\mathbf{G}'_{\mathbf{tact}}) = L(\mathbf{G}_{\mathbf{tact}})$ *for every TDES* $\mathbf{G} \in \mathbb{S}^{-1}_{a,b_1,b_2}$ *if and only if* $b_1 = b_2 = 0$, *where for each* $\mathbf{G} \in \mathbb{S}^{-1}_{a,b_1,b_2}$, $\mathbf{G}' := S^{-1}_{a,b_1,b_2}(\mathbf{G})$.

Proof: It is obvious that $\mathbf{G} \in \mathbb{S}^{-1}_{a,b_1,b_2}$ if and only if $\mathbf{G}' \in \mathbb{S}_{a,b_1,b_2}$. The result follows by a direct application of Theorem 2. ■

The result is not unexpected. It says that if the time bounds are all integer multiples of a common integer $n > 1$, then our system is to some extent overspecified. Although the original system contains more accurate timing information which may be used elsewhere, the scaled system is identical with respect to its possible event sequences.

That this result is also quite useful from a practical point of view should now be obvious. Recall that typically the TATG for a given timed discrete-event system is obtained by first constructing its TTG and then projecting out all occurrences of *tick*. Now, the state size of a TTG grows quite rapidly with an increase in *tick* frequency. (It is easy to construct an example similar to that shown in Figure 1 for which the number of states and transitions both grow as $\mathcal{O}(n^r)$, where n is the *tick* density and r is the number of self-looped events.) Using Corollary 4, we can dramatically reduce the state size of the TTG for certain TDES and hence the computational effort required in obtaining its TATG.

Composition is also a source of combinatorial explosion of TDES. We now show that scaling for the purposes of state space reduction can be applied to composed systems, before or after composition, without changing the resulting system provided all subsystems can be scaled by a common factor $n \in \mathbb{N}$. The result can also be combined with the compositional model reduction techniques described in [12, 11].

Lemma 5 *Suppose* $\mathbf{G_1}, \mathbf{G_2} \in \mathbb{S}^{-1}_n$, *for some integer* $n \in \mathbb{N}$. *Then* $\mathbf{G_1} \parallel \mathbf{G_2} \in \mathbb{S}^{-1}_n$ *and*

$$S^{-1}_n(\mathbf{G_1}) \parallel S^{-1}_n(\mathbf{G_2}) = S^{-1}_n(\mathbf{G_1} \parallel \mathbf{G_2}).$$

It should be clear that inverse scaling can only be applied in special cases. In [5], we investigate other variants of the scaling functions introduced here. Specifically, we are able to derive language containment results by reintroducing nonzero offsets b_1 and b_2. This naturally leads to approximate scaling methods. For example, in [8], the authors use the scaling function ${}_\llcorner S^{-1\urcorner}_n$ modeled after S^{-1}_n but with $l'_\sigma := \left\lfloor \frac{1}{n} l_\sigma \right\rfloor$, and $u'_\sigma := \left\lceil \frac{1}{n} u_\sigma \right\rceil$ (i.e., the floor and ceiling respectively) whenever $u_\sigma < \infty$. The function ${}_\llcorner S^{-1\urcorner}_n$ is used in order to synthesize suboptimal supervisory controllers for TDES.

4 Dense Time Limit

We now begin our study of the limiting behaviour of affine scalings as *ticks* become dense in the real time axis. In this section, we merely sketch the details of our findings. The interested reader is referred to [5] for additional details.

Suppose we are given a TDES $\mathbf{G}$ and that for each $n \in \mathbb{N}$, $\mathbf{G_n} := S_n(\mathbf{G})$. Our specific goal is to investigate whether the limit

$$\mathbf{G}' := \lim_{n\to\infty} \mathbf{G_n},$$

is meaningful. If the system $\mathbf{G}'$ can in fact be defined, then we would also like to study some of its properties. It is clear that $\mathbf{G}'$ need not be in $\mathbb{G}(\Sigma)$, solely by considering the time bounds for its timed events. Notice that if $l_\sigma > 0$ then $l'_\sigma = \infty$ in the limit, which violates the definition of TDES.

Instantaneity and Simultaneity of Events

For the purposes of the discussion below, we assume that the system $\mathbf{G}'$ as defined above exists, ignoring the above mentioned problem regarding the new definition of time bounds. We now give further evidence that such a system is not a TDES over Σ. A second problem, more subtle than the one discussed in the above paragraph, exists with $\mathbf{G}'$. Consider once again the TDES $\mathbf{G}$ whose ATG is that of Figure 1. If the activity α has time bounds $(1, 1)$, then the first occurrence of α must lie between the first and the second *tick* of the global clock. If *ticks* occur at a rate of one per unit of time in $\mathbf{G}$, then we see that the first α occurs at some time $t \in (1, 2)$. Now scale $\mathbf{G}$ by a factor of n to obtain the TDES $\mathbf{G_n}$. Under this transformation, the timed event $(\alpha, 1, 1)$ maps to (α, n, n) and so α must first occur between the n^{th} and the $(n+1)^{st}$ occurrence of *tick* in $\mathbf{G_n}$. It is not difficult to see that given any $n \in \mathbb{N}$, if the *tick* frequency increases to n *ticks* per unit time in $\mathbf{G_n}$, then the first occurrence of α in $\mathbf{G_n}$ must be at some time $t \in (1, 1 + \frac{1}{n})$. In the limit as n becomes large, this interval is empty and hence $\mathbf{G}'$ is ill-defined.

The problem we encountered above results from the interleaving semantics of TDES in which no two events are allowed to occur simultaneously. Since $tick \in \Sigma_t$, it follows that no activity $\sigma \in \Sigma$ may occur at the precise moment in time that the digital clock is updated. One solution is to declare *tick* a "special" event whose occurrence can coincide with that of any other event $\sigma \in \Sigma$. The interpretation is that *tick* is simply an external signal which denotes the passage of one unit of time. By redefining TDES in this way, the occurrence of *tick* will always precede the occurrence of σ in a string $s \in L(\mathbf{G})$ in which *tick* and $\sigma \in \Sigma$ have occurred simultaneously in the real world. Two things should be clear at this point. First, by redefining TDES in this fashion, we have not enlarged the behaviour of any given TDES $\mathbf{G}$. Second, we note that splitting hairs in this manner is rooted in the mathematical description of the model rather than physical reality since measuring time with infinite precision is impossible to accomplish. Nonetheless, we make this definition and examine its consequences. To this end, consider the example of the previous paragraph once again. In this new framework, the first occurrence of the event α in $\mathbf{G_n}$ lies in the interval $[1, 1+\frac{1}{n})$. In the limit, the interval collapses to the single point $\{1\}$ and is thus well-defined. However, note that in $\mathbf{G}'$, α has become an instantaneous event, a feature which is reserved only for *tick* in the original TDES definition.

It should also be clear that two instantaneous (non-*tick*) events can be made to occur simultaneously with this redefinition and that $\mathbf{G}'$, regardless of how it is defined, is not a TDES in the conventional sense. This motivates the introduction of a new representation for TDES which is invariant under scaling. It is believed that such a new representation will provide further insight into how the limiting system $\mathbf{G}'$ should be defined and help elucidate more of its interesting features.

TDES and Discrete Systems

Before concluding this section, we will examine one last property of TDES that can be gleaned by applying the limit as *ticks* become dense in the real line. The simultaneity encountered above is a salient feature of synchronous (viz. clock driven) systems which use interleaving semantics [4, 3, 7, 9]. We will now show that given a TDES $\mathbf{G}$, we can define a corresponding synchronous system $\mathbf{H}$ for which $L(\mathbf{G_{tact}}) = P_t(L(\mathbf{H}))$, providing some justification for the remark in [1] that "discrete-time model[s] can be viewed as a special case [of fictitious clock models] where the events happen only in lock-step with the *ticks*".

Recall the discussion following the statement of Theorem 2. It should be clear that strings of the form $s_0\, tick^n\, s_1\, tick^n s_2 \ldots (tick)^n s_k$ are sufficient to generate the TATG of $\mathbf{G_n}$. We also now know that all events in the string s_0 must occur in the time interval $[0, \frac{1}{n})$, all the events in s_1 occur in $[1, 1+\frac{1}{n})$, and so on. In the limit as n gets large, we see that all events in the string s_j must occur exactly at time $t = j$. It is now clear how one would define a synchronous transition system whose untimed language is precisely that of our limiting system. The end result is similar to the digitization of timed state sequences described in [10]. It should also be clear that creating a TDES whose untimed language is precisely that of an arbitrary synchronous system is just as easily accomplished.

5 Conclusions and Future Work

The paper introduces a three parameter family of scaling functions on timed discrete-event systems. It is shown that linearity of the scaling function is an *a priori* necessary and sufficient condition to ensure invariance of the TATG. We described inverse scaling and its role in model reduction and showed that the operation of scaling commutes with composition.

Current work in the area is focused on extending the methods of this paper to scale real-time temporal logics involving *tick* transitions. It is hoped that the cost of model-checking some TDES can be significantly reduced in this fashion. Further research will be directed toward the study of multirate TDES in which different subsystems have different *tick* frequencies and toward the introduction of a scaling invariant representation of TDES.

Acknowledgment

The first author would like to thank Peyman Gohari for many fruitful discussions regarding the subject matter of this paper.

References

[1] R. Alur, C. Courcoubetis, and D. Dill. Model-checking in dense real-time. *Information and Computation*, **104**(1), pages 2–34, 1993.

[2] R. Alur and D. Dill. A theory of timed automata. *Theoretical Computer Science*, **126**, pages 183–235, 1994.

[3] A. Benveniste, P. Le Guernic, and C. Jacquemot. Synchronous programming with events and relations: The SIGNAL language and its semantics. *Science of Computer Programming*, **16**(2), pages 103–149, 1991.

[4] G. Berry and L. Cosserat. The ESTEREL synchronous programming language and its mathematical semantics. In *CMU Seminar on Concurrency*, LNCS 197, pages 389–448, 1985.

[5] S.E. Bourdon. Design and verification of real-time control systems. PhD thesis, Systems Control Group, Department of Electrical and Computer Engineering, University of Toronto, *In Preparation.* (A full version of the current paper can also be downloaded from `http://control.toronto.edu/~bourdon`).

[6] B.A. Brandin and W.M. Wonham. The supervisory control of timed DES. *IEEE Transactions on Automatic Control*, **39**(2), pages 329–342, 1994.

[7] P. Caspi, D. Pilaud, N. Halbwachs, and J. Plaice. LUSTRE: A declarative language for programming synchronous systems. In *Proceedings of the 14th Annual Symposium on Principles of Programming Languages*, pages 178–188, 1987.

[8] P. Gohari and W.M. Wonham. Reduced supervisors for timed discrete-event systems. In R. Boel and G. Stremersch, editors, *Discrete Event Systems: Analysis and Control*, SECS 569, pages 119–130, 2000.

[9] D. Harel. Statecharts: A visual formalism for complex systems. *Science of Computer Programming*, **8**, pages 231–274, 1987.

[10] T.A. Henzinger, Z. Manna, and A. Pnueli. What good are digital clocks? In *Proceedings of the 19th International Colloquium on Automata, Languages and Programming (ICALP 1992)*, LNCS 623, pages 545–558, 1992.

[11] M. Lawford. *Model Reduction of Discrete Real-Time Systems.* PhD thesis, Systems Control Group, Department of Electrical and Computer Engineering, University of Toronto, 1997. Also available as Systems Control Group Report No. 9702.

[12] M. Lawford, J.S. Ostroff, and W.M. Wonham. Model reduction of modules for state-event temporal logics. In *Formal Description Techniques IX: Theory, Application, and Tools*, pages 263–278, 1996.

[13] P. Ramadge and W.M. Wonham. Supervisory control of a class of discrete-event processes. *SIAM Journal on Control and Optimization*, **25**(1), pages 206–230, 1987.

Modelling of Batch Process Plants by FSMs

S. Knap
Queen's University
Department of ECE
Kingston, Ontario, Canada
K7L 3N6
Email: knaps@ee.queensu.ca

T. Shepard[1]
Royal Military College
Department of ECE
Kingston, Ontario, Canada
K7K 5L0
Email: shepard@rmc.ca

M.M. Bayoumi
Queen's University
Department of ECE
Kingston, Ontario, Canada
K7L 3N6
Email: bayoumi@ee.queensu.ca

Abstract

This paper presents a novel approach to modelling of batch process plants based on finite-state machines (FSM). Our goal is to develop a model that is computationally tractable and useful for scheduling. Based on supervisory control theory for discrete-event systems (DES) [10] we propose a simple but expressive modeling approach for a class of batch production lines. One of the nice features of this approach is the fact that the composition of automata gives automatically an optimal supervisor for the system. Since this composition is optimal an additional step to find the supremal language[10] is not necessary. In addition, this approach does not require to model precedence constraints explicitly, thus greatly simplifying the modeling process. Precedence constraints are satisfied by an appropriate modeling and choice of the set of synchronization events. We prove that precedence constraints are not needed, and demonstrate that modeling of constraints by automata is difficult and not intuitive. We show in detail how the physical structures found in batch plants can be modelled for scheduling purposes. The basic types of physical structures [2] presented are single path, multi path, and network. The proposed approach is verified on clusters of different machines for both single- and multi-product cases. Extensive numerical results are provided. This modeling framework constitutes a point of departure for the main topic of our research, namely scheduling of batch processes with no intermediate storage.

1 Introduction

The idea of modeling and control design of batch process plants by automata is quite recent. A design scheme for control of a batch process plant by a timed Petri Net is proposed in [8]. In this work each activity is modelled by two transitions and one place. The first transition denotes the starting of the activity while the second denotes its end. Performance analysis of a small plant is investigated based on simulations. A similar approach is presented in [3].

In [1,5,13] Petri nets are used to model basic functions of batch plants and recipes. These authors use the Ramadge-Wonham [10] framework to provide a supervisor for a batch plant. Other models of batch plants are mostly based on mixed integer linear or nonlinear programming techniques. See for example [12,14,16,17]. In our research we use finite state machines [6] as a modelling tool. In the first stage of our research we developed untimed models of a class of batch processes. Subsequently we extended the basic untimed model to include a time variable. The timed model of a discrete-event system is presented in [7].

Analysis and modeling of DES is computationally very expensive. For this reason we have to be very rigorous and parsimonious when we model a physical system using finite automata. If the model has too many states, current technology won't allow a supervisor to be constructed for this model. Although in general our framework does not impose any limitations on the number of states in the model of the system, we make every effort to keep the number of states low in the model.

---

[1] Please send any correspondence to this author.

From the scheduling point of view we are interested in most cases in the sequencing of operations (or activities) and their start and end times. We are not concerned too much about the details of a particular activity. We only wish to know when the particular activity will start and when it will end, and which machine will be assigned to this activity. The start and end operations are assumed to be very short, or almost instantaneous. Given these requirements we come up with a simple few-state model of a component of a manufacturing system, where the term a *few-state* typically means a two- or three-state machine as shown in Figure 1.

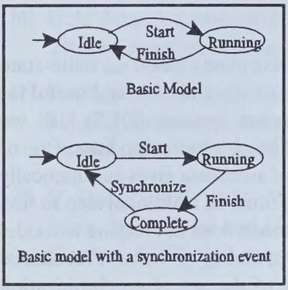


Figure 1 Untimed model of a machine.

We model a machine in the simplest case using two states *Idle* and *Running*, and two transitions *Start* and *Finish*. The labels are self-explanatory. Very often instead of descriptive labels such as *Start* or *Running* we will use *0* or *1* for brevity, and instead of *Start* or *Finish* we will use shorter labels, typically three- or four-character labels such as *Asa* or *Afa,* to denote 'machine A starts processing batch a' and 'machine A finishes processing batch a' respectively. We will usually assume, without any loss of generality, that the start events are *controllable*, and that the finish events are *uncontrollable.* For a discussion of the meaning of controllable and uncontrollable events the reader may refer to [10]. This assumption should allow us to model a large class of systems. In the case of more than one machine it will be advantageous to model some of the machines using three states and an extra transition *synchronize*, where the event *synchronize* usually denotes an event with the same label in a down-stream machine. The advantage of this approach is shown later. Here we only mention that the transition *synchronize* goes out from the state *Complete*, meaning that the machine has completed its operation, but is waiting for some synchronization event before it enters *Idle*.

If we assign time values to the *Start* and *Finish* events we will get a model which is suitable for optimum time control. It is also possible to assign some positive cost to the events in the model, thus extending the class of scheduling problems that can be solved by this approach.

From the scheduling point of view this model should be sufficient for a large class of problems. If someone is interested in detailed operations under the state *Running,* the *Running* state can be broken down into more states. These states can be seen as substates of *Running*, or *Running* can be considered as a superstate for its substates [4,11]. In practice, however, the detailed operations under *Running* are under control of a local controller whose purpose is exactly the sequencing and timing of operations on the given machine. The job of the scheduler is to initiate the operations on the machine, and the local controller takes care of the details. For this reason modelling of detailed operations under *Running* is not required for our purposes.

Below we show in detail how the physical structures found in batch plants [2] can be modelled for scheduling purposes. In this work we focus on a class of batch plants with no intermediate storage between processing units. This class of plants is frequently found in chemical batch plants, where a batch of a product is moved from one processing unit to another, for example from an evaporator to a reactor, and then to a casting machine. The basic types of physical structures presented in [2] are single path, multi path, and network. The approach taken in this paper is to demonstrate two modelling techniques, both based on FSMs, and both in the framework of the supervisory control theory.

The first technique uses the principle of 'separation of concerns', as advocated by Ramadge and Wonham [10]. In this approach a model of the system, the so-called *plant*, and a model of the constraints of the system, the so-called *specification* are two separate entities. The task of building of constraints for the system can be separated from the task of building of an unconstrained model. For example, the system to be controlled can be first modelled without any constraints, then all the constraints for the system can be specified. A reversal of these two activities is possible: model first the constraints then the system, or perform these tasks in parallel.

The second technique does not use the constraints as a separate entity. The constraints are modelled directly in the model of the system to be controlled. In this approach there is no specification automaton.

While the first technique is intuitively appealing and may seem to be more general, it presents some difficulties. The problem is that modelling of constraints by automata is not intuitive and introduces complexity that can in many cases be avoided by following our second approach. We demonstrate both techniques, while at the same time a model for a class of batch production lines is developed. FSM models were developed manually. The operations on the component FSMs such as *shuffle*, *meet*, *sync* (synchronous product), or *sup* (supremal language) [15] were generated automatically by the software that we developed during this research. The verification of correctness of our algorithms for *shuffle*, *meet*, *sync* and *sup* was done by a thorough code inspection, testing and comparison of the outputs generated by our software with the outputs generated by existing software packages [15, 9]. The comparison of the outputs between the various packages was done for small problems only.

2 Single-Path Structures

A single path structure is a group of machines through which a batch passes in a sequential manner. In this configuration the output product of one stage is used as an input product by the next stage. To this category belongs a single unit such as a reactor, or several units processing products sequentially. In our research we analysed two- and three-stage single-path production lines. Here we present a two-stage single path production line only.

A block diagram of a two-stage single path batch production line (BPL) is shown in Figure 2. The Figure also shows an untimed FSM model for this system. Figure 3 shows an FSM model of the constraints *Spec* that have to be satisfied for this configuration.

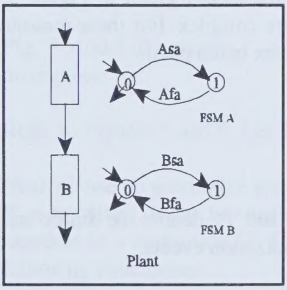


Fig. 2 FSM model of a BPL.

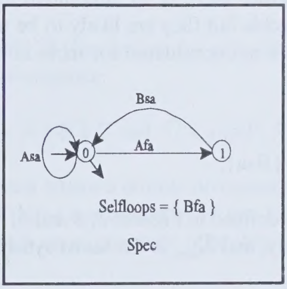


Fig. 3 Constraints for a BPL.

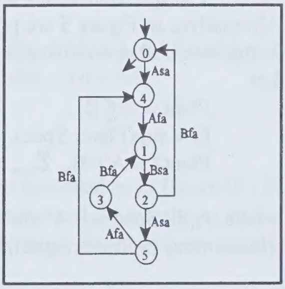


Fig. 4 Supervisor for a BPL.

In words, these constraints say that *A* has to finish first before *B* can start. Notice that the selfloop

Bfa is not shown explicitly. It is listed aside to reduce clutter. In this simple case it does not matter how the selfloops are shown, but it will matter for more complex cases. A realization of the supremal language sup(*Plant, Spec*) is shown in Figure 4.

There is nothing wrong with the model shown in Figs. 2 and 3. However, the model of the system will be much simpler if we take advantage of a set of synchronization events. In the model shown in Figure 5 the two machines share one common event, namely event *Bsa*.

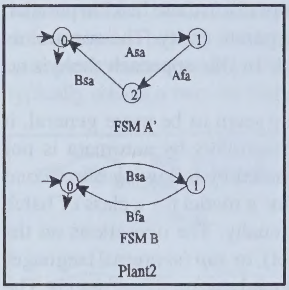


Figure 5 Model of a two-stage production line with a synchronization event.

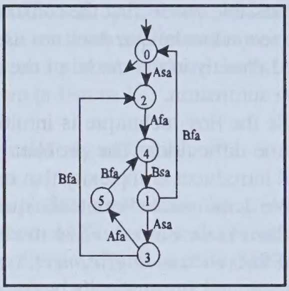


Figure 6 Synchronous product of the automata shown in Fig. 5.

It is assumed that the operation of the system shown in Figure 5 is as follows:
Machine A can start processing batch *a* (event *Asa)* at any time. Once A starts, it enters state 1. When event *Afa* (A finishes batch *a)* occurs, A enters state 2. Machine B cannot start if A is in state 0 or 1. It can only start *(Bsa)* if A is in state 2. When machine A is in state 2 it must wait until B starts. When B starts, the transitions *Bsa* in both automata occur simultaneously.

The operation of the two automata together as we have just described is exactly modelled by the synchronous product. We see that *Bsa* is a synchronization event in the sense that it couples the two machines in a particular way: A is not available again until B starts. This mechanism is very convenient as it is more intuitive and simplifies the model. Notice that we did not introduce any new event to define the synchronization. We simply used one of the pre-existing ones.

The question arises whether there is an alternative for the model given in Figure 5. Alternatives to Figure 5 are possible but they are likely to be more complex. For these reasons in this work other alternatives are not considered for more complex batch plants.

Let

Plant = A || B,
$\mathbb{C}$ = sup(Plant, Spec),
Plant2 = A'$\sqcap$B, Σ_{sync} = {Bsa},

where, A, B, Spec and A' are as defined in Figures 2, 3 and 5, '||' and '$\sqcap$' denote the shuffle and synchronous product respectively, and Σ_{sync} is the set of synchronization events.

Remark 1

Plant2 is an optimal supervisor for a two-stage single-path production line, i.e, Plant2 = sup(Plant, Spec).

The sync product of FSMs given in Fig. 5 is shown in Fig. 6. It can be noticed that it is the same as the model of the supervisor shown in Figure 4, with an appropriate reordering of labels.

It turns out that a similar result holds for any number of stages of a single path batch production line.

3 Multi-Path Structures

A simple multi-path arrangement of units consists of single-path production lines working in parallel independently, without any transfer between the lines [2]. In this case each single path can be modelled as demonstrated in the previous section. If, however, two or more single-path production lines share some of the machines, additional constraints have to be added to the model of a single-path production line. Below we analyse the basic constructs that are necessary to model multi-path configurations.

3.1 Path Divergence

In this section we show how a double divergence (or split) is modelled.
Double divergence is a configuration in which a section of a single path is followed by a section of two single paths. In this arrangement there is a section where one stage of the line (a machine) is followed by two machines that work in parallel. A FSM model of a double split is given in Figures 7 (plant) and 8 (constraints).

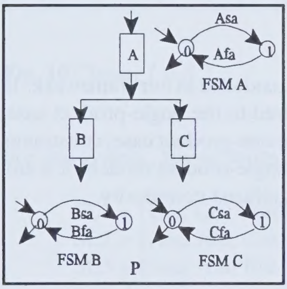


Fig. 7 Model of a double divergence.

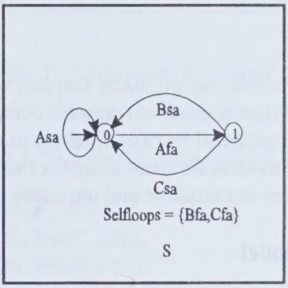


Fig. 8 Constraints for a double divergence.

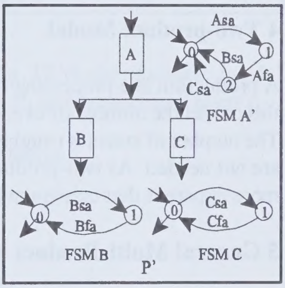


Fig. 9 Model of a double split with sync events.

Refer to Figures 7 and 8. Let $P = A \parallel B \parallel C$ and $\mathbb{C} = \sup(P, S)$.

Next we use an alternative approach where a double divergence is modelled as in Figure 9. Let $P' = A' \sqcap B \sqcap C$. It can be verified that a realization of P' has 12 states and 24 transitions and is identical to a realization of $\mathbb{C}$ with an appropriate relabeling of states. From this we get the following conclusion:

Remark 2
P' is an optimal supervisor for *P*, i.e., P' = sup(P, S).

A similar result holds for any degree of divergence.

3.2 Path Convergence

Double convergence is a configuration in which a section of two single paths is followed by a section of a single path. In this arrangement there is a section where one stage of the line (two parallel machines) is followed by a single machine.

Our model of a double convergence is similar to the model of a double divergence. Similarly to the case of double divergence an optimal supervisor can be constructed directly from the model of the plant, without explicit constraints. A similar result holds for any degree of convergence.

3.3 Network Structures

A network structure is a multiple path arrangement that has one or more loops. In this configuration the batch may revisit a given unit one or more times before it reaches its final processing stage. Modelling of a network configuration by FSMs is basically the same as in the case of single- and multi-path architectures. Again, modelling of constraints by automata is not required. More importantly, constraint automata again introduce complexity and are error prone.

4 Two-product Model

A production line processing two different products can also be modelled in our framework. In this model the number of events (and transitions) doubles compared to the single-product case. The number of states is roughly multiplied by two. Similarly to the one-product case, constraints are not needed. As two-product models are more complex than single-product models, it is still more apparent that constraints are not intuitive and introduce significant complexity.

5 General Multi-Product Model

Modelling of production lines processing more than two products is similar to modelling of the two-product cases. One can notice that one product added to the model increases the number of states of the model by one, and the number of transitions by two. As the number of products increases modelling using a graphical representation of the machines and the required constraints become very cumbersome. For this reason, in the case of more than two products the model must be implemented directly in software. We use this approach below when we analyse clusters of machines.

6 Cluster Analysis

A generic component of a batch production line can be modelled as a cluster composed of a number of units working on three levels. On level one there is one unit. On level 2 there may be a number of units working in parallel. On level three there is one unit. If level two had one unit only, then we would have a single-path production line. In the next sections we analyse a cluster with two parallel units at level 2.

When we model multi-stage lines we will use the following notation: a-b-c-...-n.p to denote a k-stage production line with *a* units at stage 1, *b* units at stage 2, *c* units at stage 3, and *n* units at stage *k*. The .p refers to the number of products being modelled. For example, a 1-2-1.3 production line is a line with one unit at stage 1, two units at stage 2, and one unit at stage 3; this production line may process up to three different product types.

6.1 Cluster 1-2-1.1

Figures 10 and 11 show a cluster of machines with two parallel units without synchronization events.

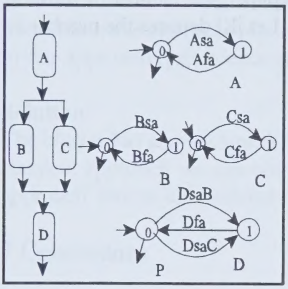

A
Asa
Afa
0
1
A
B
C
Bsa
Bfa
0
1
B
Csa
Cfa
0
1
C
D
DsaB
Dfa
DsaC
0
1
P
D

Fig. 10 Cluster 1-2-1.1.

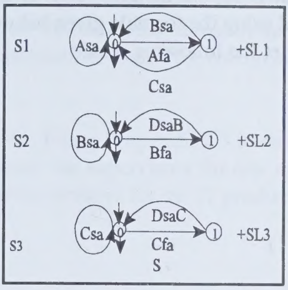

S1
Asa
Bsa
Afa
0
1
Csa
+SL1
S2
Bsa
DsaB
Bfa
0
1
+SL2
S3
Csa
DsaC
Cfa
0
1
+SL3
S

Fig. 11 Constraints for cluster 1-2-1.1.

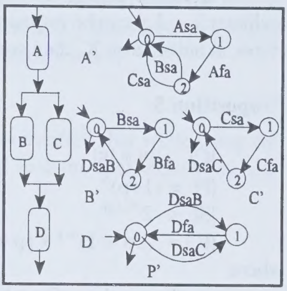

A
A'
Asa
Bsa
Csa
Afa
0
1
2
B
C
Bsa
DsaB
Bfa
B'
Csa
DsaC
Cfa
C'
D
DsaB
Dfa
DsaC
0
1
P'
D

Fig. 12 Model of cluster 1-2-1.1 with synchronization events.

For this system , the selfloops, SL1, SL2 and SL3 are defined as follows:

$SL1 = \{Bfa, Cfa, DsaB, DsaC, Dfa\}$,
$SL2 = \{Asa, Afa, Csa, Cfa, DsaC, Dfa\}$,
$SL3 = \{Asa, Afa, Bsa, Bfa, DsaB, Dfa\}$,

and, $P = A \parallel B \parallel C \parallel D$,
$S = S_1 \cap S_2 \cap S_3$,
$\mathbb{C} = \sup(P, S)$,

The same system can be modelled as shown in Figure 12, where four synchronization events are used, namely, *Bsa, Csa, DsaB and DsaC.*

This system is is defined by: $P' = A' \sqcap B' \sqcap C' \sqcap D$. It turns out that the following is true:

Remark 3
P' is an optimal supervisor for P, i.e., $P' = \sup(P, S)$.

Remark 4
This technique extends to any arbitrary number of machines at stage 2, and any arbitrary number

of products.

6.2 Cluster Analysis - Multi-Product Case

The classical modelling and analyses presented above, the plant plus spec technique, is appropriate for small problems only. As problems become larger, this approach becomes quickly tedious, and simply unmanageable. For example, cluster 1-4-1.2 requires 10 constraint automata with very long list of selfloops. Analysis of larger problems than this one with the classical approach and tools such as TCT [15] or OTCT [9] cannot be done. By using our modelling approach with synchronization events and our n-way algorithm for constructing synchronous products we can compute significantly larger problems (with more products) mechanically, without specifying the constraints explicitly.

If the modelling technique is used as demonstrated in this paper, the number of states in a cluster 1-x-1.p can be estimated using the formula given below. Let |K| denotes the number of states in automaton K, then we get the following result:

Proposition 5

$$|\mathbb{C}| \leq |P| * |S|$$
$$|P| = (1+p)^{N}$$
$$|S| \leq 2^{(N-1)p}$$
$$|P'| = (2p+1)^{N-1} \times (p+1)$$

where

p - the number of products, $p > 0$,
N - the number of machines in the cluster, $N > 2$.

The table below summarizes the results for Clusters 1-4-1 up to ten products. It shows the number of events, states and transitions in the synchronous product as well as the time required to build the synchronous product. The computations were performed on a Sun Enterprise 3500 machine running at 400MHz and equipped with 3Gbytes of RAM. These results were generated by the software developed in this research.

# of Products	# of Events	# of States	# of Transitions	Time [sec]
1	15	486	1,755	0
2	30	9,375	36,750	1
3	45	67,228	268,569	6
4	60	295,245	1,186,812	28
5	75	966,306	3,893,175	90
6	90	2,599,051	10,479,690	244
7	105	6,075,000	24,499,125	561
8	120	12,778,713	51,527,544	1,208

9	135	24,760,990	99,819,027	2,478
10	150	44,925,111	181,052,550	5,234

Table 1 Results for cluster 1-4-1.

The example given below shows that the technique based on the set of synchronization events allows us to solve significantly larger problems than the plant-plus-spec approach.

Example

Consider Cluster 1-4-1.p modelled as demonstrated in the previous sections, and assume the we have a computer that can store supervisors that have less than, say, 10^8 states. Find the largest p for which one can find a supervisor for this cluster using: a) the classical (plant+spec) approach, b) our approach (precedence constraints modelled by synchronization events).

Solution:

The cluster has six machines (N=6). Using Proposition 5 it can be verified that while using the classical approach we can construct the supervisors for one and two-product cases only, our approach lets us to construct the supervisors for up 11 products.

7 Conclusions

The idea of modelling batch process plants by automata is, to our knowledge, about 10 years old. However, most of the authors use Petri nets as a modelling tool [1,3,5,8,13]. In this paper we developed a modelling approach based on FSMs in the framework of supervisory control theory. Since one of the main problems of this theory is computational complexity, we made every effort to develop a model that has a minimal number of states and transitions and at the same time captures useful properties of the system being modelled. Our goal has been to develop a model that is computationally tractable and useful for scheduling. For this reason we considered only start and finish events, with automata that have two or three states per product. We demonstrated that the plant-plus-spec modelling approach in the form suggested by the basic supervisory control theory is not appropriate for the class of problems we considered in this work.

First, modelling of constraints by automata is difficult and not intuitive. In our opinion this is one of the main obstacles for a wide industrial acceptance of the supervisory control theory.

Second, constraint automata introduce complexity. In order to generate a specification automaton it is required to take an intersection of the constraint automata. This operation usually generates an automaton with the number of states equal to the product of the number of states of the composite automata. The size of this automaton can be very large. In addition, to construct a supervisor, a synchronous product of the plant P and the spec S is usually constructed. If the algorithm for constructing the sync product requires a table of size $|P| * |S|$ then larger problems cannot be solved by this approach.

The results presented in this paper indicate that the construction of a supervisor by the use of our approach for a cluster of three, four, five and six machines is computationally feasible given the current status of the computer technology. In the case of one-, two- and three-products the execution times for all configurations were almost negligible. Our approach allowed us to construct supervisors for systems significantly larger than would have been possible by the use of the classical approach.

References

[1] K. Akesson. Recipe Coordination in Chemical Batch Processes. Thesis for the Degree of Licentiate of Engineering, Chalmers University of Technology, Goteborg, Sweden, 1999.

[2] American National Standards Institute/Instrument Society of America. Batch Control. Part 1: Models and Terminology, 1995.

[3] T. Gu and P.A. Bahri. Timed Petri-Net Representation for Short Term Scheduling of Multiproduct Batch Plants, Proceedings of the ACC, San Diego, California, June 1999, pp.4092-4096.

[4] D. Harel. Statecharts: A visual formalism for complex systems, Sci. Comput. Prog., Vol. 8, pp.231-274, 1987.

[5] A. Hellgren. Modelling and Implementation Aspects of Supervisory Control. Thesis for the Degree of Licentiate of Engineering, Department of Signals and Systems, Chalmers University of Technology, Goteborg, Sweden, 2000.

[6] J. Hopcroft and J. Ullman. Introduction to Automata Theory, Languages, and Computation, Addison-Wesley Publishing Company, 1979.

[7] S.Knap, T.Shepard, M.M.Bayoumi. Modification to the Brandin-Wonham Timed Model of DESs, 38th Annual Allerton Conference on Communication, Control, and Computing, Oct. 2000.

[8] S. Lloyd and Y. Md. Salleh. Modeling and Control Design of Batch Process Plant b y Timed Petri Net, Proceedings of the 30th CDC, 1991, pp. 1531-1536.

[9] S.D. O'Young. Object TCT: User's Guide, System Control Group, Department of Electrical Engineering, University of Toronto, Toronto, Report No. 9312, September, 1993.

[10] P. Ramadge, W. Wonham. The Control of Discrete Event Systems, Proceedings of the IEEE, Vol. 77, No. 1, January, 1989, pp. 81-97.

[11] Bran Selic, Garth Gullekson and Paul T. Ward. Real-Time Object-Oriented Modeling, John Wiley & Sons, Inc., 1994.

[12] S. Subrahmanyam, G.K. Kudva. M.H. Bassett, and J.F. Pekny. Application of Distributed Computing to Batch Plant Design and Scheduling, AIChE Journal, Vol. 42, No 6, pp. 1648-1661, June 1996.

[13] M. Titus. Control Synthesis for Batch Processes, Ph.D Thesis, Chalmers University of Technology, Gothenburg, Sweden, 1995.

[14] V.T. Voudouris and I.E. Grossmann. MILP Models for Scheduling and Design of a Special Class of Multipurpose Batch Plants, Computers Chem. Engng Vol. 20, No. 11, pp. 1335-1360, 1996.

[15] W.M. Wonham. DES Lecture Notes and TCT/TTCT software package. University of Toronto,Toronto,1999, available from: http://www.control.toronto.edu/people/profs/wonham.html.

[16] M.C. Wellons and G.V. Reklaitis. Scheduling of Multipurpose Batch Chemical Plants. 1. Formulation of Single-Product Campaigns, Ind. Eng. Chem. Res, Vol. 30, No 4, pp. 671-688, 1991.

[17] E.C. Yamalidou, E.P. Patsidou and J.C. Kantor. Modeling Discrete-Event Dynamical Systems for Chemical Process Control - A survey of Several New Techniques, Computers chem. Engng, Vol. 14, No. 3, pp. 281-299, 1990.

Modification to the Brandin-Wonham Timed Model of DESs

S. Knap
Queen's University
Department of ECE
Kingston, Ontario, Canada
K7L 3N6
Email: knaps@ee.queensu.ca

T. Shepard[1]
Royal Military College
Department of ECE
Kingston, Ontario, Canada
K7K 5L0
Email: shepard@rmc.ca

M.M. Bayoumi
Queen's University
Department of ECE
Kingston, Ontario, Canada
K7L 3N6
Email: bayoumi@ee.queensu.ca

Abstract

In this paper we present a modification to the timed model of discrete-event systems (DESs) proposed by Brandin and Wonham in [2] In the Brandin-Wonham model (BW) each event has lower and upper time bounds within which it must occur. The model is very appealing as it allows to model both delays and deadlines. An unfortunate 'feature' of this model is the clock tick which is part of the event set Σ. The fact that the clock tick is treated as an event has serious repercussions on the complexity of the model. Although the BW model is very attractive, it can only be used for very small problems [11]. The problem with this model is a state space blowup as a function of timing constraints. Our timed model of DESs does not use the tick and is more tractable. We modify the BW model by removing the tick and providing a simple mechanism to advance the timers. This model is applied to modelling and control of batch production lines. The new model allows us to solve significantly larger problems. Extensive numerical results are provided.

1 Introduction

The supervisory control theory introduced by Ramadge and Wonham [12] is based on basic untimed automata, and as such it does not include any notion of time. Untimed automata model the sequencing of events but not the times at which the events take place. Although events occur in time, there is no time variable in this model. In [2,3,4,5] Brandin and Wonham extend the untimed model of DESs to a timed model of DESs. In this model each event $\sigma \in \Sigma$ has lower and upper time bounds (l_σ, u_σ) within which it must occur. The model is very appealing as it allows to model both delays and hard deadlines. An unfortunate 'feature' of this model is the clock tick which is added to the event set Σ. In this model, as time passes by, a global clock ticks causing event tick to occur. The fact that the clock tick is treated as an event has serious repercussions on the complexity of the model.

In this paper we present a timed model that does not use the tick. The model was strongly motivated by the timed model proposed in [2]. Although the BW model is very attractive, it can only be used for very small problems. The problem with the BW model is a state space blowup as a function of timing constraints. To demonstrate this problem consider Example 1.

Example 1
Figure 1 shows a FSM model of a two-stage production line. Two machines A and B are processing parts of product *a*. Product *a* is first processed by machine A then by machine B.

---

[1] Please send any correspondence to this author.

Transfer time between the two machines is added to the processing time of machine A and is not modelled explicitly. The meaning of event labels is the following:

Asa, (Bsa) - machine A (machine B)starts processing product a,
Afa, (Bfa) - machine A (machine B) finishes processing product a,

Assuming that the time bounds for the events are Asa(0, ∞), Afa(1,1), Bsa(0,∞), Bfa(1,1), the BW timed models of A and B are given in Figure 2. Their timed composition *P* is shown in Figure 3. *P* has 10 states and 15 transitions.

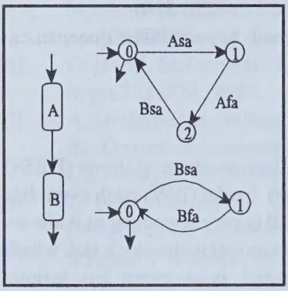


Fig. 1 Simple production line.

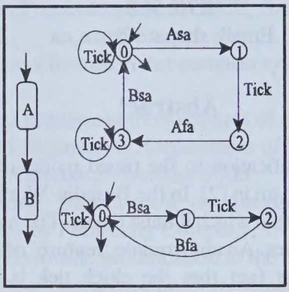


Fig. 2 BW timed graphs for the system shown in Fig. 1.

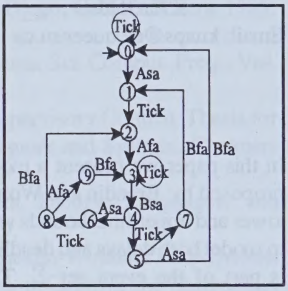


Fig. 3 Composition of A and B shown in Fig. 2.

This seems to be a nice solution. What is then the problem ? In this example the time bounds for the *Afa* and *Bfa* events are set to one unit only. What would be the size of P if we increased the time bounds for *Afa* and *Bfa* to 10, 100, 1000 units accordingly ? Table 1 gives the answer. The results given in Tables 1 to 3 were calculated by construction using TTCT [13], and/or OTCT [10], and our own software.

Bounds for Afa and Bfa	Timed A [States/Transitions]	Timed B [States/Transitions]	Composition of A and B [States/Transitions]
1,1	4/6	3/4	10/15
10,10	13/15	12/13	91/105
100,100	103/105	102/103	5,356/5,460
1000,1000	1,003/1,006	1,002/1,004	503,506/505,514

Table 1 Number of states/transitions for different time bounds for the system given in Fig. 2.

Clearly this model can only be used for problems with small time bounds.

In the example given above the start events *Asa* and *Bsa* have their time bounds very large, namely (0,∞). This basically means that these events can occur at some instant *t* after they become enabled. The instant *t* is not specified. In particular, they may occur right away (lower bound 0), or they may not occur at all (upper bound equal to ∞). This modelling feature seems to be very attractive as it allows to model the events that may, or may not occur at all. However, the price we pay for this feature is very high. To see this, consider Example 2.

Example 2
Consider the system given in Figure 1. Assume now that the start events *Asa* and *Bsa* occur right away after their enablement, i.e., *Asa*(0,0) and *Bsa*(0,0). The finish events remain unchanged. Let us recompute Table 1 for this model. Table 2 shows the results.

Bounds for Afa and Bfa	Timed A [States/Transitions]	Timed B [States/Transitions]	Composition of A and B [States/Transitions]
1,1	4/4	3/3	8/9
10,10	13/13	12/12	26/27
100,100	103/103	102/102	206/207
1000,1000	1,003/1,004	1,002/1,003	2,006/2,010

Table 2 Results for Example 2.

By comparing Table 1 and 2 one can see that the size (number of states/transitions) of the timed models of A and B is almost the same in both examples respectively. This result suggests that in the BW model the fact of setting the upper bound of a transition to infinity has little effect on the size of the model. However, this seems to be only true for the models of the component automata. A dramatic difference in sizes is seen in their respective compositions. By contrast, as it can be verified later, our time models of A, B and their composition for Example 2 have 3/3, 2/2, and 7/8 states/transitions respectively four all four cases.

The lower and upper bounds of an event in the BW model provide a mechanism to model the uncertainty of the time instants at which events occur. Since in general, it is very difficult to predict precisely the time of occurrence of any event the BW approach seems to be very realistic. Unfortunately, this feature turns out to be very expensive. This is demonstrated in Example 3.

Example 3
Consider the same system as in Example 1. Assume that Asa(0,0), Bsa(0,0). Find the BW model for a) Afa(1,10), Bfa(1,10), b) Afa(1,100), Bfa(1,100), c) Afa(1,1000), Bfa(1,1000).

The results are given in Table 3.

Bounds for Afa and Bfa	Timed A [States/Transitions]	Timed B [States/Transitions]	Composition of A and B [States/Transitions]
1,1	4/4	3/3	8/9
1,10	13/22	12/21	35/72
1,100	103/202	102/201	305/702
1,1000	1,003/2,002	1,002/2,001	3,005/7,002

Table 3 Results for Example 3.

2 Our Model of Timed Discrete-Event Systems

In this section we present our timed model of DES. The model is based on the timed model proposed by Brandin and Wonham in [2]. We removed the tick and modified the mechanism for adjusting the timers. The fact that we do not model the tick allows us to solve significantly larger problems. The framework, other concepts and terminology such as the definitions of enabled and eligible events are the same or similar as in the BW model. In particular, the event set is composed of controllable and uncontrollable events, and *forcible* events are those that can preempt other eligible events.

Similarly to BW model we introduce time to the untimed model G_u by associating the events with the time bounds in which they must occur. However, the time bounds do not have to be the natural numbers. Similarly to [AD90] we use continuous time. By convention time 0 is assigned to the first event. Later events are assigned to nonnegative real numbers from $\mathbb{R}$. However, in practice we are interested in the delays between consecutive events. For this reason, in a realization of this model we associate a timer with each event. Initially the timers are set to 0. As events occur the timers of all enabled events are incremented by the time interval Δ_{min} that will have elapsed between the current event and the next event that will occur. We do not decrement timers at each tick of a 'global' clock, and thus, do not generate unnecessary states and transitions. This is a fundamental difference between the BW model and our model. In the BW model the automaton remains in a given state only one unit of time in general. In our model the automaton remains in a given state until an event occurs.

Similarly to the BW model we define a timed discrete-event system G

$$G = (Q, \Sigma, \delta, q_0, Q_m),$$

where the components of G are similar to the components of G in the BW model. However, the clock tick is not part of the event set Σ.

If we associate with event $\sigma \in \Sigma$ the time of its occurrence t_σ then the pairs (σ, t_σ) are called *timed events*. The timer interval for σ, denoted by Γ_σ, is defined by $\Gamma_\sigma = [0, t_\sigma]$.

With each event σ we associate a timer T_σ whose role is to keep track of the passing time. Thus T_σ can have any value in the interval $[0, t_\sigma]$.

If the event set Σ is composed of events *a*, *b*, *c*, ..., *k* then the state space Q in the timed model is given by

$$Q = Q_u \times \Gamma_a \times \Gamma_b \times \Gamma_c \times ... \times \Gamma_k .$$

Intuitively a state *q* in Q is a pair consisting of an untimed state and the current timer values.

The last expression can be written in a more compact form:

$$Q = Q_u \times \prod \Gamma_\sigma, \sigma \in \Sigma, \text{ where } \prod \text{ denotes the cross product operation.}$$

The initial state of timed DES is q_o with the timers of all events from Σ set to 0. The marker state of the timed system is Q_{um} with some arbitrary timer values. The transition function δ defines a transition from state q_1 to state q_2 in the timed model. Below we use the notation "$q_2 = \delta(q_1, \sigma)!$" to denote "transition from q_1 to q_2 triggered by σ is defined".

In our research we consider three cases: (i) deterministic case, i.e., for each $\sigma \in \Sigma$ $l_\sigma = u_\sigma$, (ii) for each $\sigma \in \Sigma$ $l_\sigma \le u_\sigma < \infty$, and (iii) for each $\sigma \in \Sigma$ $l_\sigma \le u_\sigma$, with a possibility of $u_\sigma = \infty$. Due to the space limitation in this paper we provide only a treatment of the deterministic case (i).

2.1 Deterministic Case

In this case we assume that the time instants at which events occur are known precisely, i.e.,for each $\sigma \in \Sigma$ $l_\sigma = u_\sigma$. The logic underlying this case is simpler and the model is smaller than in the case of unequal time bounds. Nevertheless, when we relax the requirement for equal time bounds it will turn out that the algorithm for generation of a timed model in the first case (Algorithm 2) requires only a minor modification to generate a timed model in a general case.

Definition 1 (Transition in a timed model)
A transition in the timed system is defined if and only if the corresponding transition in the untimed model is defined and the timer associated with the event triggering this transition has expired.

A timer T_σ expires when its value reaches l_σ. Using the notation from the automata theory this definition can be written in the following form:

Let $q_2 = \delta(q_1, \sigma)$, $q_2, q_1 \in Q$,
then $\delta(q_1, \sigma)!$ if and only if $\delta_u(q_{u1}, \sigma)!$ and $T_\sigma = l_\sigma$.

Notice that:

$$q_1 = (q_{u1}, x),\ q_2 = (q_{u2}, x), \text{ where } q_{u1}, q_{u2} \in Q_u, \text{ and } x \in \prod \Gamma_\sigma,\ \sigma \in \Sigma.$$

Thus, in the timed model transitions are defined for eligible events only.

Below we describe an algorithm for creating the list of eligible events at state q.

Algorithm 1 (for creating the list of eligible events at state q)
Input : Untimed state q, timers, time bounds
Output: eligibleEvents,
1. Create the list of enabled events *enabledEvents*
2. Find an event that will fire first, i.e., for each event in *enabledEvents*, compute the current delay Δ_σ, where $\Delta_\sigma = \max(0, l_\sigma - T_\sigma)$, and select the smallest Δ_σ; denote it as $\Delta_{min.}$
3. Create the list of eligible events *eligibleEvents*, i.e, select those events from *enabledEvents* whose delay $\Delta_\sigma = \Delta_{min}$, assign *delay* Δ_{min} to each eligible event.

While in the BW the timers are decremented by one at each tick of the clock, in our model the timers are incremented[1] by Δ_{min}.

Initially, at the initial state all the timers are reset to 0. Next, Algorithm 1 is executed. Algorithm 1 returns the list of eligible events - the list of events whose timers have expired. We assume that one of these events will actually occur and that will determine the next state, but in the model we take into account all eligible events. When event σ occurs the timers are adjusted as specified below.

---

[1] An implementation of our algorithm in which the timers are decremented is possible. However, we found that incrementing the timers is more natural.

Let for each $\tau \in \Sigma$ $T_{1\tau}$ and $T_{2\tau}$ be the timer values for τ in states q_1 and q_2 respectively.
Let $\sigma, \tau \in \Sigma$ then $q_{u2} = \delta_u(q_{u1}, \sigma)$, and

if $\tau \neq \sigma$, $T_{2\tau} := 0$ if not $\delta_u(q_{u2}, \tau)!$, (timer value for event τ is reset to 0),
$T_{2\tau} := T_{1\tau}$ if $\delta_u(q_{u2}, \tau)!$, (timer value for event τ is unchanged)

if $\tau = \sigma$, $T_{2\tau} := 0$, (timer value for event σ is reset to 0).

In words, when an event σ occurs its timer is reset to 0. The timers of other events are also reset to 0 except those timers whose events are defined in the new state; in the latter case the values of the timers are not changed.

Below we give a formal algorithm for constructing a timed graph for our model. The essence of the algorithm is the mechanism in which we look ahead to create a list of the events that will occur next. Thus creating new states for each tick is not necessary. This approach reduces the number of states and transitions significantly.

Algorithm 2 (for constructing a timed automaton)

```
Input: untimed automaton plus time bounds for each event.
Output: timed automaton defined by two arrays, states, and transitions
0. Set all timers to 0, create $q_0$, add $q_0$ to currentLevel and to the states tables,
1. For each state in currentLevel
        1.1 Execute Algorithm 1
        1.2 For each eligible event
                1.2.1 Set timers to the timer values of the current state,
                1.2.2 Increment the timers of all enabled events by the delay of the current event,
                      i.e, for each $\sigma$ in enabledEvents: $T_\sigma = T_\sigma + delay$,
                1.2.3 Calculate the state newState to which FSM will enter when this event occurs,
                1.2.4 If newState does not exist in states add it to the states and nextLevel tables,
                1.2.5 Create a transition t from the current state to newState, assign delay to t,
                      add t to transitions.
2. If (nextLevel is empty) goto 3
   else currentLevel = nextLevel, goto 1
3. End
```

Notice that in our timed model with each transition a delay is associated, i.e., the amount of time that must elapse before the transition is triggered. The delay is measured with respect to the time of entrance to the current state.

A similar idea for adjusting timers in a timed automaton is presented in [6, Ch. 5]. However, one of the key differences between this algorithm and the one presented in [6] is the fact that our model allows for more than one outgoing transition from a state, while their model allows for one transition only. There, if at a state there are two or more eligible events (two or more timers have expired on enabled events) only one event is selected, according to some priority rule. This approach has serious repercussions on optimal control as it eliminates potential optimal paths from the graph.

Remark 1
The number of states generated by Algorithm 2 is bounded by $O(m \prod |t_\sigma|)$, where m denotes the number of states in the untimed model.

Example 4
This example illustrates how Algorithm 2 works. It also shows the main difference between the BW and our model. Fig. 4 shows an untimed and timed model generated by our Algorithm. The timed events are as follows: A(2,2), B(4,4) and C(5,5). Fig. 5 shows the BW timed model for this system.

Notice that in Fig. 4 we show the timer values at each state in the format (x-a.b.c), where x is a state in the untimed model and a, b, c are current values of T_a, T_b, T_c. Also, each transition has a label followed by the delay that must elapse before the transition is taken. While our model has four states and five transitions, the BW model has nine states and ten transitions. Clearly, in this case there is no advantage to having tick states in the model. They introduce complexity only and can be removed without any lost of information.

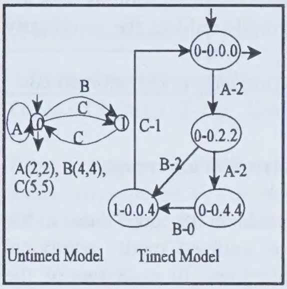


Fig. 4 Model for Example 4.

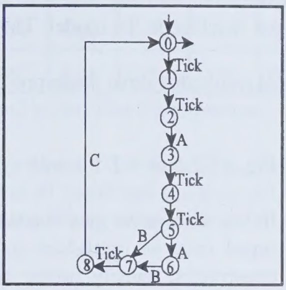


Fig. 5 BW model of the system shown in Fig. 4.

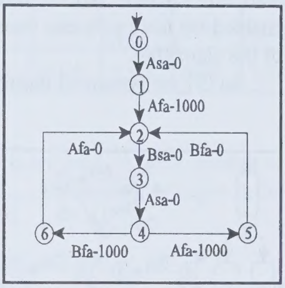


Fig.6 Composition of A and B for Example 5.

Example 5
The results for Example 2 are given in Table 2. By contrast, our time models of A, B and their composition for Example 2 have 3/3, 2/2, and 7/8 states/transitions respectively four all four cases. In Figure 6 we show our timed model of the composition of A and B for the case where Afa(1000,1000) and Bfa(1000,1000). In this case the BW composition of A and B has 2,006 states and 2,010 transitions.

3 Search Algorithms

The implementation of Step 1.2.4 in Algorithm 2 turns out to determine whether the whole algorithm can be judged as 'fast' or 'slow'. Here is a description of the problem. As new states are generated and added to the *states* table, it becomes larger and larger. Whenever a new state is added a search must be performed to determine whether this is an existing state or not. For small problems, the states table can be scanned and each element of the table can be compared with the new state. When the number of states reaches a few thousands of elements this approach is not acceptable. We see that there is a need for a smart search algorithm. Several approaches are possible, even in the family of binary search algorithms. The results can be really surprising. We investigated several different algorithms on a cluster of 4, 5 and 6 machines for different number of products. The best results were obtained by the use of the hash table approach. This search technique is known to give a constant average time for performing insertions, deletions and finds of elements of a set [7].

4 Numerical Results

The timed models were tested on clusters of three, four, five and six machines. The machines were modelled as two- or three-state automata as described in [9]. A remarkable feature of the modelling approach taken in [9] is the fact that the synchronous product of the component automata yields an optimal supervisor for the system. The notation '1-x-1.p' used below denotes a three-stage production line with one machine at stage one, x machines at stage 2 and 1 machine at stage 3; *p* denotes the number of products being processed by the cluster. An example of a model of a cluster 1-2-1.1 is shown in Figure 6. A Sun Enterprise 3500 machine running at 400MHz equipped with 3Gbytes of RAM was used in this study.

In the general model of a timed discrete-event system we assumed that every event has a timer associated with it. Notice, however, that assuming that start events occur as soon as they are enabled we need only one timer par machine in the model. This greatly reduces the complexity of the algorithm.

In [9] we presented numerical results for three-stage production lines in the untimed case.

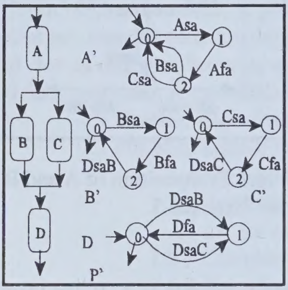


Fig. 6 Cluster 1-2-1.1 with synchronization events.

In this section we give similar results in the timed case. In the timed case we introduce to the untimed model additional constraints, namely timing constraints. In each run of the algorithm we assume the following constraints:

a) unit at level 1 requires 20 minutes to process any batch,
b) all units at level 2 require 100 minutes to process any batch,
c) unit at level 3 requires 30 minutes to process any batch,
d) transfer between stages occurs in zero time,
e) start events are not delayed,
f) all start events are controllable and forcible, while finish events are uncontrollable.

Table 5 gives the results for the most complex case, namely cluster 1-4-1.p, for up to **7** products. In each case a timed model of the synchronous product of the component machines was constructed by Algorithm 2. The column *Time* gives the time required by the algorithm to construct the timed model of the cluster. The column '# of Possible States' displays the upper bound for the number of states in the timed model. Notice that the actual number of states is a tiny fraction of the possible number of states.

# of Products	# of Events	# of Possible States	# of States	# of Transitions	Time [sec]
1	15	8.539×10^{7}	843	998	0
2	30	5.465×10^{9}	33,501	43,868	3
3	45	6.225×10^{10}	322,255	441,810	25
4	60	3.497×10^{11}	1,656,153	2,331,992	129

5	75	1.334×10^{12}	5,976,531	8,564,030	464
6	90	3.984×10^{12}	17,180,533	24,925,428	1,364
7	105	1.005×10^{13}	42,140,631	61,703,978	3,654

Table 5 Results for Cluster 1-4-1 Timed Case.

Table 6 gives the ratio of the number of states (transitions) in the timed model to the number of states (transitions) in the untimed model for cluster 1-2-1.

Num. of Products	1	2	3	4	5	6	7	8	9	10
States [Timed/Untimed]	1.02	1.41	1.62	1.75	1.84	1.9	1.94	1.98	2.01	2.03
Trans. [Timed/Untimed]	0.47	0.65	0.76	0.84	0.89	0.92	0.95	0.98	0.99	1.01

Table 6 Cluster 1-2-1.p. Untimed versus timed comparison.

Table 7 compares the BW model with ours. The following symbols are used: S_1 - the number of states in our model, T_1 - the number of transitions in our model, S_{BW} - the number of states in the BW model, T_{BW} - the number of transitions in the BW model, and

$$\Delta s = ((S_{BW} - S_1)/S_1) * 100,$$
$$\Delta t = ((T_{BW} - T_1)/T_1) * 100.$$

# of Products	S1	T1	S_{BW}	T_{BW}	Δs	Δt
1	55	60	533	538	869	797
2	529	656	5,497	5,624	939	757
3	2,227	2,928	23,545	24,246	957	728
4	6,385	8,688	68,033	70,336	966	710
5	14,671	20,420	157,021	162,770	970	697
6	29,185	41,280	313,273	325,368	973	688
7	52,459	75,096	564,257	586,894	976	682
8	87,457	126,368	942,145	981,056	977	676
9	137,575	200,268	1,483,813	1,546,506	979	672
10	206,641	302,640	2,230,841	2,326,840	980	669

Table 7 BW versus ours. Cluster 1-2-1.

5 Conclusions

The BW model has appealing features as it models both time delays and hard deadlines. In addition it provides control mechanisms through disablement and forcing. Although the BW model is very attractive, it can only be used for very small problems. The problem with the BW model is a state space blowup as a function of timing constraints. Our timed model of DESs does not use the tick and is more tractable. We modify the BW model by removing the tick and providing a simple mechanism to advance the timers. This simple approach allowed us to solve significantly larger problems.

The size of a timed model is usually greater than the size of its untimed couterpart. In the case of a model with the tick there may be a very large explosion of the state space. In the model without the tick there is usually an increase in the size of the model, but the increase is not so great.

In this paper we provide only a treatment of the case when delays are known precisely. This requirement may in many cases be difficult to satisfy. It turns out that Algorithms 1 and 2 require only minor modification to accommodate the case of unequal time bounds.

References

[1] R.Alur and D.Dill. Automata for modeling real-time systems. In Lecture Notes in Computer Science 443, editor, Proc. of the 17th International Colloquium on Automata and Programming. Springer-Verlag, 1990.

[2] B.A. Brandin and W.M. Wonham. Supervisory Control of Timed Discrete-Event Systems, IEEE Transactions on Automatic Control, Vol. 39, No. 2, February 1994, pp. 329-341.

[3] B.A. Brandin, W.M. Wonham and B. Benhabib. Manufacturing Cell Supervisory Control - A Timed Discrete Event System Approach, Proceedings of the 1992 IEEE International Conference on Robotics and Automation, Nice, France, May 1992, pp. 931-936.

[4] B.A. Brandin, W.M. Wonham and B. Benhabib. Modular Supervisory Control of Timed Discrete Event Systems, 1992 Allerton Communication and Control Conference, Urbana-Champaign, September 1992, pp. 624-632.

[5] B.A. Brandin, W.M. Wonham and B. Benhabib. Manufacturing Cell Supervisory Control - A Modular Timed Discrete Event System Approach, Proceedings of IEEE International Conference on Robotics and Automation, 1993, pp. 846-851.

[6] C. Cassandras and S. Lafortune. Introduction to Discrete Event Systems. Kluwer Academic Publishers, 1999.

[7] A. Drozdek. Data Structures and Algorithms in C++, PWS Publishing Company, 1996.

[8] R. Karp and M. Held. Finite-State Processes and Dynamic Programming, SIAM J. Applied Mathematics, Vol. 15, No. 3, May 1967, pp. 693-718.

[9] S. Knap, T. Shepard, M.Bayoumi. Modelling of Batch Process Plants by FSMs, 38th Annual Allerton Conference on Communication, Control, and Computing, Oct. 2000.

[10] S.D. O'Young. Object TCT: User's Guide, System Control Group, Department of Electrical Engineering, University of Toronto, Toronto, Report No. 9312, September, 1993.

[11] K. Rudie, N. Shimkin and S. O'Young. Timed Discrete Event Systems: A Manufacturing Application, Proceedings of the 1994 Conference on Information Sciences and Systems, Princeton, NJ, March 16-18, 1994, pp. 374-381.

[12] P. Ramadge, W. Wonham. The Control of Discrete Event Systems, Proceedings of the IEEE, Vol. 77, No. 1, January, 1989, pp. 81-97.

[13] W.M. Wonham. DES. Lecture Notes and TTCT Software Package. University of Toronto, Toronto,1999, available from the web site http://www.control.toronto.edu/people/profs/wonham.html.

An Adaptive Filter Robust to Data Uncertainties*

Tareq Y. Al-Naffouri[†] and Ali H. Sayed[‡]

Abstract

This paper considers the problem of adaptive filtering in the presence of uncertainties in the regression data. A recursive procedure is derived that is based on solving local optimization problems that attempt to alleviate the worst-case effect of data uncertainties on filter performance. The resulting procedure turns out to have similarities with leakage-based adaptive filters.

1 Introduction

In real applications there are always discrepancies between an assumed model and the actual data. Such discrepancies or data uncertainties can be due to several factors such as unmodeled dynamics, disturbances, simplifications during the modeling phase, or even finite precision effects. In this paper, we focus on uncertainties in the regression data in the context of adaptive filtering. In particular, we suggest a procedure that attempts to alleviate the worst-case effect of the uncertainties on filter performance. The resulting algorithm includes leakage and has connections with leaky-LMS, as we shall explain later.

Thus consider a reference sequence $\{d(k)\}$ that arises from a linear model of the form

$$d(k) = \boldsymbol{x}_k^T \boldsymbol{w}_o + v(k), \tag{1}$$

where $v(k)$ accounts for measurement noise, $\boldsymbol{x}_k$ is a regression vector (taken as a column vector), and $\boldsymbol{w}_o$ is an unknown column vector of size M that we wish to estimate. Adaptive algorithms for estimating $\boldsymbol{w}_o$ from the data $\{\boldsymbol{x}_k, d(k)\}$ rely on recursive procedures that attempt to match the output of the adaptive filter to the reference signal $\{d(k)\}$ in some optimal manner.

In this work, we examine the effect of one source of uncertainty that can interfere with the performance of an adaptive filter. It pertains to the fact that at any time instant, k, the regression vector $\boldsymbol{x}_k$ may not be known perfectly but only to within a certain uncertainty. In order to indicate this fact explicitly, we replace the linear model (1) by the alternative expression

$$d(k) = (\boldsymbol{x}_k + \delta \boldsymbol{x}_k)^T \boldsymbol{w}_o + v(k) \tag{2}$$

---

*This work was supported in part by a grant from the National Science Foundation under Award number ECS-9820765.

[†] T. Y. Al-Naffouri is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305. E-mail: nafouri@stanford.edu. URL: www.stanford.edu/~nafouri

[‡] A. H. Sayed is with the Electrical Engineering Department, University of California, Los Angeles, CA 90095. E-mail: sayed@ee.ucla.edu. URL: www.ee.ucla.edu/asl

where $\delta \boldsymbol{x}_k$ denotes the uncertainty in $\boldsymbol{x}_k$. In this context, $\boldsymbol{x}_k$ represents the *nominal* value of the actual regression vector and $\delta \boldsymbol{x}_k$ represents the unknown disturbance to $\boldsymbol{x}_k$.

Motivated by the formulation in [1] for design problems with uncertainties (see also [2]–[4] for related applications), we consider two models for the uncertainty $\{\delta \boldsymbol{x}_k\}$:

1. Uncertainty in factored form. In this first model we assume that the uncertainty lies along a certain direction, i.e., it satisfies a relation of the form
$$\delta \boldsymbol{x}_k \; = \; a \boldsymbol{E}_a \quad \text{for all } k, \tag{3}$$
where $\boldsymbol{E}_a$ is a known column vector and a is an unknown arbitrary scalar bounded by unity, $|a| \leq 1$. Different choices for $\boldsymbol{E}_a$ would correspond to different modeling assumptions on $\delta \boldsymbol{x}_k$. For example, the choice $\boldsymbol{E}_a = \text{col}\{\eta, \eta, \ldots, \eta\}$ means that each entry of $\delta \boldsymbol{x}_k$ can be regarded as bounded by η, which is useful to model finite precision effects on the data. Similarly, the choice $\boldsymbol{E}_a = \text{col}\{\eta, 0, \ldots, 0\}$, or similar sparse vectors $\boldsymbol{E}_a$, allow us to model the effect of impulsive disturbances. Also, a column vector $\boldsymbol{E}_a$ with distinct entries allows us to handle different levels of uncertainties in the various entries of $\boldsymbol{x}_k$.

2. Bounded uncertainty. In this second model we assume that the uncertainty is simply bounded, say
$$\|\delta \boldsymbol{x}_k\| \leq \eta, \tag{4}$$
for some known scalar η and for all k. Here the notation $\| \cdot \|$ denotes the Euclidean norm of its argument.

2 Adaptive Filter Design

Given the uncertain model (2), we can proceed to formulate an optimization problem that seeks to estimate the weight vector $\boldsymbol{w}_o$ in some robust manner. By a robust procedure we mean one that attempts to alleviate the effect of the uncertainties on filter performance, especially under worst-case conditions.

Thus assume that we already have a weight-vector estimate at time k, say $\boldsymbol{w}_k$, and let us seek an estimate for time $k+1$, say $\boldsymbol{w}_{k+1}$. We shall obtain $\boldsymbol{w}_{k+1}$ from $\boldsymbol{w}_k$ by solving the following optimization (min-max) problem:
$$\min_{\boldsymbol{w}_{k+1}} \max_{\delta \boldsymbol{x}_k} J(\boldsymbol{w}_{k+1}) \tag{5}$$
where the cost function $J(\boldsymbol{w}_{k+1})$ is quadratic in $\boldsymbol{w}_{k+1}$ and is defined by
$$J(\boldsymbol{w}_{k+1}) = |\; d(k) \; - \; (\boldsymbol{x}_k + \delta \boldsymbol{x}_k)^T \, \boldsymbol{w}_{k+1} \;|^2 \; + \; \alpha \|\boldsymbol{w}_{k+1} - \boldsymbol{w}_k\|^2, \tag{6}$$
and α is a positive regularization parameter. The maximization in (5) is carried out over any assumed model for the uncertainty $\{\delta \boldsymbol{x}_k\}$, such as the ones indicated by (3) or (4).

The choice of the cost function (6) is motivated by the fact that in the absence of uncertainties, it is known that the classical normalized LMS algorithm can be motivated by regarding it as the solution to a regularized least-squares problem of this kind (see, e.g., [6]). Here, instead of simply minimizing a quadratic cost function, we are minimizing its worst-possible residual in the presence of the uncertainties (which corresponds to solving a min-max problem). To solve this problem we rely on the least-squares estimation theory with uncertain data developed in [1] — see also [2]–[4] and [5] for related discussions and applications.

To proceed, we introduce the quantities,

$$\begin{aligned} \boldsymbol{w} &\triangleq \boldsymbol{w}_{k+1} - \boldsymbol{w}_k \\ e(k) &\triangleq d(k) - \boldsymbol{x}_k^T \boldsymbol{w}_k, \end{aligned}$$

which are defined in terms of the given data $\{\boldsymbol{x}_k, d(k)\}$ and the weight estimates $\{\boldsymbol{w}_k, \boldsymbol{w}_{k+1}\}$, as well as the following quantity, which incorporates the effect of the uncertainty $\delta\boldsymbol{x}_k$:

$$u(k) \triangleq \delta\boldsymbol{x}_k^T(\boldsymbol{w} + \boldsymbol{w}_k). \tag{7}$$

It then follows that we can rewrite the expression for $J(\boldsymbol{w}_{k+1})$ as a function of $\boldsymbol{w}$,

$$J(\boldsymbol{w}) \;=\; |\boldsymbol{x}_k^T \boldsymbol{w} - e(k) + u(k)|^2 \;+\; \alpha\|\boldsymbol{w}\|^2, \tag{8}$$

where $u(k)$ can be interpreted as representing the aggregate model for the uncertainties.

Using (3), we can replace (5) by the optimization problem

$$\min_{\boldsymbol{w}} \max_{|u(k)| \leq \phi(\boldsymbol{w})} J(\boldsymbol{w}), \tag{9}$$

where the function $\phi(\boldsymbol{w})$ is defined by

$$\phi(\boldsymbol{w}) = \|\boldsymbol{E}_a^T(\boldsymbol{w} + \boldsymbol{w}_k)\|, \tag{10}$$

and $\{\boldsymbol{w}_k, \boldsymbol{E}_a\}$ are given. Using (4) instead, we obtain the same optimization problem (9) with the following alternative expression for $\phi(\boldsymbol{w})$:

$$\phi(\boldsymbol{w}) = \eta\|\boldsymbol{w} + \boldsymbol{w}_k\|. \tag{11}$$

Observe that we can regard this second form for $\phi(\boldsymbol{w})$ as a special case of (10) by assuming $\boldsymbol{E}_a = \eta I$. This observation allows us to treat both cases (10) and (11) simultaneously in the sequel: in one case $\boldsymbol{E}_a$ is a column vector and in the other a multiple of the identity matrix.

3 Solving the Optimization Problem

We solve the optimization problem (9) by using the solution method of [1]. To begin with, we rewrite the min-max problem (9) in the following equivalent min-min form:

$$\min_{\lambda \geq 1} \min_{\boldsymbol{w}} \left[C(\boldsymbol{w}, \lambda) + \alpha\|\boldsymbol{w}\|^2 \right], \tag{12}$$

where the two-variable function $C(\boldsymbol{w}, \lambda)$ is given by

$$C(\boldsymbol{w}, \lambda) = \left(1 + (\lambda - 1)^\dagger\right)\left(\boldsymbol{x}_k^T \boldsymbol{w} - e(k)\right)^2 + \lambda\phi^2(\boldsymbol{w}), \tag{13}$$

and

$$(\lambda-1)^{\dagger}=\begin{cases} 0 & \text{if } \lambda=1 \\ \frac{1}{\lambda-1} & \text{otherwise.} \end{cases}$$

The parameter λ is a positive scalar.

We can now search for the minimum over $\boldsymbol{w}$ of (12) by setting the derivative of the cost function, $C(\boldsymbol{w},\lambda)+\alpha\|\boldsymbol{w}\|^2$, with respect to $\boldsymbol{w}$ equal to zero. This yields the equation

$$\left(\alpha I+\left(1+(\lambda-1)^{\dagger}\right)\boldsymbol{x}_k\boldsymbol{x}_k^T\right)\boldsymbol{w}+\frac{1}{2}\lambda\nabla\phi^2(\boldsymbol{w})=\left(1+(\lambda-1)^{\dagger}\right)\boldsymbol{x}_k e(k), \tag{14}$$

where the notation $\nabla\phi^2(\boldsymbol{w})$ denotes the gradient of $\phi^2(\boldsymbol{w})$ with respect to $\boldsymbol{w}$. We should stress that the optimum (and unique) $\boldsymbol{w}^o$ that results from solving (14) is a function of λ and, accordingly, should be explicitly written as $\boldsymbol{w}^o(\lambda)$. The optimum $\boldsymbol{w}^o$ over the permissible range of λ $(\lambda\geq 1)$ can then be obtained by solving the (one-dimensional) optimization problem:

$$\min_{\lambda\geq 1} G(\lambda) \tag{15}$$

where $G(\lambda)$ is defined by

$$G(\lambda)=C(\boldsymbol{w}^o(\lambda),\lambda)+\alpha\|\boldsymbol{w}^o(\lambda)\|^2. \tag{16}$$

We now invoke the special form of $\phi(\boldsymbol{w})$ to carry out the above calculations. As indicated earlier, $\phi(\boldsymbol{w})$ has the general form:

$$\phi(\boldsymbol{w})=\|\boldsymbol{E}_a^T(\boldsymbol{w}+\boldsymbol{w}_k)\|, \tag{17}$$

where $\boldsymbol{E}_a$ is either a column vector or $\boldsymbol{E}_a=\eta I$. The following calculations are valid regardless of what form $\boldsymbol{E}_a$ takes. In particular, we have

$$\nabla\phi^2(\boldsymbol{w}) \;=\; 2\boldsymbol{E}_a\boldsymbol{E}_a^T(\boldsymbol{w}+\boldsymbol{w}_k). \tag{18}$$

Upon substituting (18) into (14), and solving for $\boldsymbol{w}^o(\lambda)$, we obtain

$$\boldsymbol{w}^o(\lambda)=\left[\alpha I+\left(1+(\lambda-1)^{\dagger}\right)\boldsymbol{x}_k\boldsymbol{x}_k^T+\lambda\boldsymbol{E}_a\boldsymbol{E}_a^T\right]^{-1}\left(\left(1+(\lambda-1)^{\dagger}\right)\boldsymbol{x}_k e(k)-\lambda\boldsymbol{E}_a\boldsymbol{E}_a^T\boldsymbol{w}_k\right) \tag{19}$$

Notice that no use has yet been made of the special structure of our problem. In particular, notice that the only matrix that appears in (19) is the identity, and every other parameter is otherwise a vector or a scalar. This fact can be exploited to get a more explicit expression for $\boldsymbol{w}^o(\lambda)$ since we can invoke the matrix inversion lemma to evaluate the inverse in (19). As such, at this point, we have to treat the two classes of uncertainties we mentioned before separately.

4 The Case of Bounded Uncertainties

For bounded uncertainties (where we regard $\boldsymbol{E}_a=\eta I$), we find by the matrix inversion lemma that

$$\boldsymbol{w}^o(\lambda)=-\gamma\boldsymbol{w}_k+\lambda\frac{\gamma\boldsymbol{x}_k^T\boldsymbol{w}_k+e(k)}{\beta(\lambda-1)+\lambda\|\boldsymbol{x}_k\|^2}\boldsymbol{x}_k \tag{20}$$

where

$$\beta \triangleq \alpha + \eta^2\lambda \quad \text{and} \quad \gamma \triangleq \eta^2\lambda/\beta.$$

Or, since $\boldsymbol{w}^o(\lambda) = \boldsymbol{w}_{k+1} - \boldsymbol{w}_k$ and $e(k) = d(k) - \boldsymbol{x}_k^T\boldsymbol{w}_k$, we get

$$\begin{aligned} \boldsymbol{w}_{k+1} &= (1-\gamma)\boldsymbol{w}_k + \lambda\frac{d(k) - (1-\gamma)\boldsymbol{x}_k^T\boldsymbol{w}_k}{\beta(\lambda-1) + \lambda\|\boldsymbol{x}_k\|^2}\,\boldsymbol{x}_k & (21) \\ &= (1-\gamma)\boldsymbol{w}_k + \lambda\frac{e^w(k)}{\beta(\lambda-1) + \lambda\|\boldsymbol{x}_k\|^2}\,\boldsymbol{x}_k, & (22) \end{aligned}$$

where the modified error $e^w(k)$ is defined by

$$e^w(k) \triangleq d(k) - (1-\gamma)\boldsymbol{x}_k^T\boldsymbol{w}_k. \tag{23}$$

Four remarks are in place here:

1. Equation (22) gives the update equation for $\boldsymbol{w}_{k+1}$ in terms of a parameter λ, whose optimal value we still need to determine. Note that the values of $\{\beta, \gamma\}$ also depend on λ. As pointed out earlier (see (15)-(16)), the optimum λ is obtained by minimizing the function $G(\lambda)$ which in the present case is given by

$$G(\lambda) = \begin{cases} \frac{1}{\alpha+\eta^2}\left(\frac{\left(-\alpha e(k) - \eta^2 d(k)\right)^2}{\alpha+\eta^2+\|\boldsymbol{x}_k\|^2} + \alpha\eta^2\|\boldsymbol{w}_k\|^2\right) & \text{if } \lambda = 1 \\ \frac{\lambda}{\beta}\left[\frac{\left(\alpha\boldsymbol{x}_k^T\boldsymbol{w}_k - \beta d(k)\right)^2}{\beta(\lambda-1)+\lambda\|\boldsymbol{x}_k\|^2} + \alpha\eta^2\|\boldsymbol{w}_k\|^2\right] & \text{otherwise.} \end{cases} \tag{24}$$

2. Minimizing $G(\lambda)$ requires that we find the roots of its derivative, $dG(\lambda)/d\lambda$, or, equivalently, the roots of a fourth-order polynomial. While this could be done numerically and in closed form as well, we are more interested in deriving a simplified algorithm. For this purpose, we have observed through repeated simulations that the minimum value of the function $G(\lambda)$ generally occurs at a value that is very close to 1. In this case we can simplify the update recursions to:

$$\boldsymbol{w}_{k+1} = (1-\gamma)\boldsymbol{w}_k \; + \; \frac{d(k) - (1-\gamma)\boldsymbol{x}_k^T\boldsymbol{w}_k}{\alpha + \eta^2 + \|\boldsymbol{x}_k\|^2}\,\boldsymbol{x}_k, \tag{25}$$

   where

$$\gamma = \frac{\eta^2}{\alpha + \eta^2} \tag{26}$$

   Table 1 summarizes the adaptive algorithm for the bounded uncertainty case, where we have also incorporated a step-size parameter μ.

3. By inspecting (25) we notice that it has a form similar to leaky-LMS, except that the leakage factor $(1-\gamma)$ is applied to both terms on the right-hand side. We shall comment further on this fact below.

Table 1: *Adaptive algorithm for bounded regressor uncertainty.*

DESCRIPTION	EQUATION
Adaptation Equation	$\boldsymbol{w}_{k+1} = (1-\gamma)\boldsymbol{w}_k \; + \; \mu \, \frac{d(k)-(1-\gamma)\boldsymbol{x}_k^T\boldsymbol{w}_k}{\alpha+\eta^2+\|\boldsymbol{x}_k\|^2} \, \boldsymbol{x}_k$
Filtering Equation	$e^w(k) = d(k) - (1-\gamma)\boldsymbol{x}_k^T\boldsymbol{w}_k$
Leakage Parameter	$\gamma = \frac{\eta^2}{\alpha+\eta^2}$
Step-size	μ
Apriori Information	α (regularization parameter) η (uncertainty parameter)

4. In our development, we could have obtained similar (and more sophisticated) results with an additional uncertainty model on the additive noise $v(k)$ (see (2)). In particular, if we assume that $|v(k)| \leq \eta_v$ for some η_v, then the function $\phi(\boldsymbol{w})$ would have taken the form (compare with (17)):

$$\phi(\boldsymbol{w}) = \|\boldsymbol{E}_a^T(\boldsymbol{w}+\boldsymbol{w}_k)\| + \eta_v. \tag{27}$$

The adaptation equation we arrived at (namely (25)) remains essentially the same.

5 The Case of Factored Uncertainties

Consider (19) again for the case of factored uncertainties. Building on our observation in the second remark above, we set λ in (19) to 1 to get

$$\boldsymbol{w}^o(1) = \left[\alpha I + \boldsymbol{x}_k\boldsymbol{x}_k^T + \boldsymbol{E}_a\boldsymbol{E}_a^T\right]^{-1}\left(e(k) - \boldsymbol{E}_a\boldsymbol{E}_a^T\boldsymbol{w}_k\right). \tag{28}$$

We need to employ the matrix inversion lemma twice this time to evaluate the inverse in (28). We find after tedious but straightforward manipulations that the adaptation equation is given this time by

$$\boldsymbol{w}_{k+1} = \left\{ I - \frac{1}{\Delta}\left[\|\boldsymbol{x}_k\|^2\left(I - \frac{\boldsymbol{x}_k\boldsymbol{x}_k^T}{\|\boldsymbol{x}_k\|^2}\right) + \alpha I\right]\boldsymbol{E}_a\boldsymbol{E}_a^T\right\}\boldsymbol{w}_k + \frac{1}{\Delta}\left[\|\boldsymbol{E}_a\|^2\left(I - \frac{\boldsymbol{E}_a\boldsymbol{E}_a^T}{\|\boldsymbol{E}_a\|^2}\right) + \alpha I\right] e(k)\boldsymbol{x}_k, \tag{29}$$

where

$$\Delta \triangleq \left(\alpha + \|\boldsymbol{x}_k\|^2\right)\left(\alpha + \|\boldsymbol{E}_a\|^2\right) - \left(\boldsymbol{E}_a^T\boldsymbol{x}_k\right)^2 \tag{30}$$

6 A Comment on Leakage in Adaptive Filtering

Leakage in LMS is usually employed when the input is noisy or not persistently exciting since it helps avoid the drift problem of LMS. In this case, the adaptation equation is of the form

$$\boldsymbol{w}_{k+1} = (1-\mu\alpha)\boldsymbol{w}_k + \mu\boldsymbol{x}_k e(k), \tag{31}$$

for some α, with $e(k) = d(k) - \boldsymbol{x}_k^T\boldsymbol{w}_k$ and $d(k) = \boldsymbol{x}_k^T\boldsymbol{w}_o + v(k)$.[1] The leakage factor $(1-\mu\alpha)$ applies to the factor $\boldsymbol{w}_k$ in (31) only but not to the $\boldsymbol{w}_k$ that appears in the output error $e(k)$. This distinction relates to the bias problem that is characteristic of leaky-LMS (see also [6, 7] and [8]).

To examine this point further, let us investigate the bias problem when we introduce a leakage factor in both terms of the update equation (31), i.e., let us consider a recursion of the form

$$\boldsymbol{w}_{k+1} = (1-\mu\alpha)\boldsymbol{w}_k + \mu\boldsymbol{x}_k e^w(k), \tag{32}$$

where

$$e^w(k) = d(k) - (1-\alpha)\boldsymbol{x}_k^T\boldsymbol{w}_k. \tag{33}$$

It is easy to verify that this modified algorithm leads to unbiased weight vector estimates for i.i.d. input signals. To this end, we rewrite (32) in terms of the weight error vector, $\tilde{\boldsymbol{w}}_k = \boldsymbol{w}_k - \boldsymbol{w}_o$, by subtracting $\boldsymbol{w}_o$ from both sides of (32):

$$\begin{aligned} \tilde{\boldsymbol{w}}_{k+1} &= (1-\mu\alpha)\tilde{\boldsymbol{w}}_k - \mu\alpha w_o + \mu\boldsymbol{x}_k e^w(k) \\ &= (1-\mu\alpha)\tilde{\boldsymbol{w}}_k + \mu\left(\boldsymbol{x}_k e^w(k) - \alpha\boldsymbol{w}_o\right). \end{aligned} \tag{34}$$

Now the error $e^w(k)$ too can be written in terms of the weight error vector:

$$\begin{aligned} e^w(k) &= d(k) - (1-\alpha)\boldsymbol{x}_k^T\boldsymbol{w}_k & (35) \\ &= \alpha\boldsymbol{x}_k^T\boldsymbol{w}_o - (1-\alpha)\boldsymbol{x}_k^T\tilde{\boldsymbol{w}}_k + v(k), & (36) \end{aligned}$$

so that (34) becomes

$$\tilde{\boldsymbol{w}}_{k+1} = \left(1 - \mu\alpha - \mu(1-\alpha)\boldsymbol{x}_k\boldsymbol{x}_k^T\right)\tilde{\boldsymbol{w}}_k + \mu\alpha(\boldsymbol{x}_k\boldsymbol{x}_k^T - I)\boldsymbol{w}_o + \mu\boldsymbol{x}_k v(k). \tag{37}$$

Assuming the input to be i.i.d., of unit variance (i.e., $E\boldsymbol{x}_k\boldsymbol{x}_k^T = I$), and independent of the noise, we get upon averaging the last equation that

$$E[\tilde{\boldsymbol{w}}_{k+1}] = (1 - \mu\alpha - \mu(1-\alpha))\, E[\tilde{\boldsymbol{w}}_k],$$

a homogeneous recursion in $E[\tilde{\boldsymbol{w}}_k]$. Thus, for a small enough step-size, $E[\tilde{\boldsymbol{w}}_k]$ will converge to zero and $\boldsymbol{w}_k$ will be an unbiased estimate of $\boldsymbol{w}_o$.

But what if the input signal is correlated? We can treat the general input case by introducing the "whitening" factor $1/\|\boldsymbol{x}_k\|^2$ in the adaptation equation (32), which then becomes

$$\boldsymbol{w}_{k+1} = (1-\mu\alpha)\boldsymbol{w}_k + \mu\boldsymbol{x}_k\,\frac{e^w(k)}{\|\boldsymbol{x}_k\|^2}. \tag{38}$$

---

[1] In this section we are ignoring uncertainties in the regression vectors.

This "whitening" factor is usually introduced in the LMS algorithm to improve its convergence for a spectrally correlated input. When introduced for the leaky-LMS version (32), this factor serves the additional purpose of helping "ensure" unbiasedness for a correlated input. This gives rise to a hybrid algorithm that is composed of leaky-LMS and NLMS. The question is whether it is possible to justify these modifications and constructions in a more "rigorous" manner.

As can be seen from the discussions in the earlier sections, it turns out that these modifications are similar to what we get when we derive an adaptive algorithm in the presence of uncertainties in the regression data.

7 Simulation Results

In order to demonstrate some of the points raised in the previous section, we compared the performance of the algorithm of Table 1 with the following related leaky-algorithms:

1. The leaky-LMS algorithm.

$$\boldsymbol{w}_{k+1} = (1 - \mu\alpha)\boldsymbol{w}_k + \mu e(k)\boldsymbol{x}_k$$

2. The leaky-NLMS algorithm.

$$\boldsymbol{w}_{k+1} = (1 - \mu\alpha)\boldsymbol{w}_k + \mu\frac{e(k)}{\beta + \|\boldsymbol{x}_k\|^2}\boldsymbol{x}_k$$

In both cases, we have $e(k) = d(k) - \boldsymbol{x}_k^T\boldsymbol{w}_k$.

To ensure a fair comparison, we chose the step size so that all three algorithms would have the same convergence speed. For the same reason, the three algorithms also shared the same leakage factor (i.e., the factor multiplying $\boldsymbol{w}_k$ on the right-hand side). The algorithms were employed in identifying a four tap FIR filter. Both the nominal input and the associated uncertainty were chosen to be uniformly distributed such that

$$\frac{\max\|\delta\boldsymbol{x}_k\|}{\max\|\boldsymbol{x}_k\|} = \frac{0.4}{4} = 10\%.$$

This means that $\eta = .4$ and $\gamma = .096$. The experiment was repeated 10,000 times and the resulting learning curves are shown in Fig. 1. The top curve refers to leaky-LMS, the middle curve to leaky-NLMS, and the bottom curve to the robust-NLMS version of Table 1.

8 Conclusion

In this paper we derived an adaptive algorithm that attempts to alleviate the worst-case of the uncertainties in the regression data on filter performance. Such situations can arise, for example, in channel tracking applications for multicarrier modulation. The resulting algorithm turns out to include leakage, albeit in a manner that applies both to the weight vector $\boldsymbol{w}_k$ and to the estimation error $e^w(k)$, as shown in Table 1.

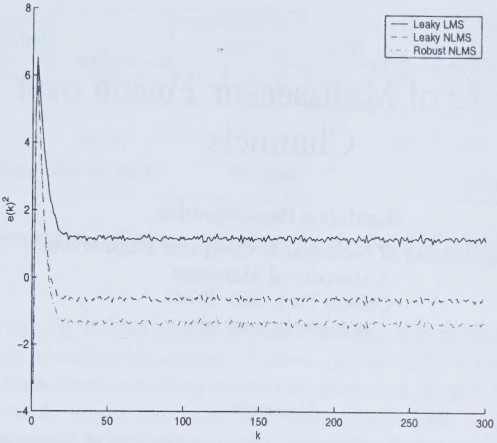


Figure 1: *Learning curves for leaky*-LMS, *leaky*-NLMS, *and the robust version of Table 1.*

References

[1] A. H. Sayed and V. H. Nascimento, "Design criteria for uncertain models with structured and unstructured uncertainties," in *Robustness in Identification and Control*, A. Garulli, A. Tesi, and A. Vicino, editors, vol. 245 of Lecture Notes in Control and Information Sciences, pp. 159–173, Springer Verlag, London, 1999.

[2] A. H. Sayed, V. H. Nascimento, and S. Chandrasekaran, "Estimation and control with bounded data uncertainties," *Linear Algebra and Its Applications*, vol. 284, pp. 259–306, Nov. 1998.

[3] A. H. Sayed and S. Chandrasekaran, "Parameter estimation with multiple sources and levels of uncertainties," *IEEE Transactions on Signal Processing*, vol. 48, no. 3, pp. 680–692, Mar. 2000.

[4] A. H. Sayed, T. Y. Al-Naffouri, and T. Kailath, "Robust estimation for uncertain models in a data fusion scenario," *Proc. IFAC System Identification Symposium,* Santa Barbara, CA, June 2000.

[5] L. E. Ghaoui and H. Hebret, "Robust solutions to least-squares problems with uncertain data," *SIAM J. Matrix Anal. Appl.*, vol. 18, pp. 1035–1064, 1997.

[6] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 3rd edition, NJ, 1996.

[7] B. Widrow and S. D. Stearns, *Adaptive Signal Processing*, Prentice Hall, NJ, 1985.

[8] V. H. Nascimento and A. H. Sayed, "Unbiased and stable leakage-based adaptive filters," *IEEE Transactions on Signal Processing*, vol. 47, no. 12, pp. 3261–3276, Dec. 1999.

Some Aspects of Multisensor Fusion over Fading Channels

Haralabos Papadopoulos
Department of Electrical & Computer Engineering
University of Maryland
College Park, MD 20742
babis@eng.umd.edu

Abstract

We consider a problem that involves noisy observations of the same static signal, collected by a network sensors, that are sequentially encoded by means of quantizer bias control and communicated over fading channels to a host for fusion. We present methods for optimizing the sensor encodings and develop maximum likelihood estimators for data fusion at the host for feedback-based and random quantizer bias control encoders. We also present computationally efficient fusion methods that exploit linear combining of the received encodings followed by appropriate processing.

1 Introduction

In this paper we examine the problem of signal estimation from sequentially encoded noisy measurements corrupted by fading. This problem arises in the context of a number of wireless sensor networks whereby each sensor obtains a noisy sequence of signal observations and must communicate these measurements efficiently and with minimal delay over a fading channel to a host.

Inherent limitations in the sensor apparatus and the modulation circuitry often dictate communication via a finite set of signal levels over the fading channel. Hence, there is a need to design methods for encoding the sensor measurements into a digitized stream prior to communication over the fading channel.

We focus on sequential digital encoding of the sensor measurements via quantizer bias control [1], whereby a properly designed control signal is added to each measurement prior to scalar quantization. The block diagram corresponding to the ith sensor and the fading channel model is shown in Fig. 1, where $A[n]$ denotes the information-bearing signal, $v_i[n]$ represents sensor noise, $s_i[n]$ denotes the sensor measurement sequence, $d_i[n]$ is the control input used at the ith sensor, $y_i[n]$ denotes the corresponding encoding sequence, α_i captures the effects of fading, $w_i[n]$ represents receiver thermal noise, and $z_i[n]$ denotes the faded encoding received by the host that is to be used to obtain a signal estimate $\hat{A}[n]$. In this paper we focus on the static signal case, *i.e.*, the case $A[n] = A$, and whereby each sensor produces a binary-valued encoding sequence $y_i[n]$. The objective is to design the sensor encoders and the associated fusion rule from the faded encodings at the host so as to minimize the mean-square estimation error (MSE) of the host signal estimate.

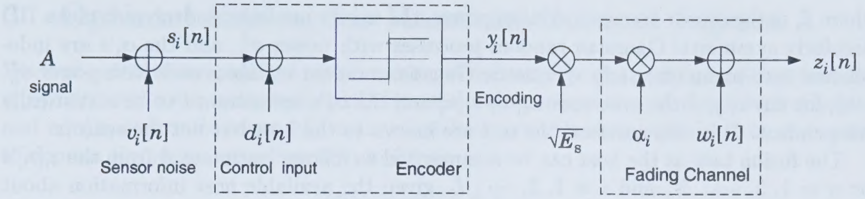


Figure 1: Communication of sensor encodings in the form of quantizer bias control over a fading channel.

This problem can be viewed as one of lossy encoding of a noisy source [2] achieved via a class of refinable encoding strategies. Various (nonrefinable) lossy encoding methods for noisy sources have been examined in the literature; see [3] and the references therein.

In this paper we focus on two important cases which differ in the amount of control input information available to the host and the freedom in the control input selection. In the first case, referred to as the random control case, we assume that the control inputs $d_i[n]$ are selected as zero-mean independent IID Gaussian random processes with power σ_d^2, and assume that the host can only exploit knowledge of the statistical characterization of the $d_i[n]$'s. This type of control signal is often exploited in lossy compression and is commonly referred to as nonsubtractive dithering; it has been shown to provide compression-distortion improvements in image and audio coding [4]. We show that there is a channel-SNR dependent optimal control power level which, when used in conjunction with a suitably designed fusion rule at the host, minimizes the host MSE. The second case, referred to as the feedback control case, includes a broadcasting channel exploited by the host to relay back to the sensors a control input adjustment sequence $d_{\text{host}}[n]$ that is computed at the host from all past received encodings. As we show, proper selection of the feedback sequence $d_{\text{host}}[n]$ allows suitably designed fusion methods at the host to come within a small loss of the encoder-free performance.

2 System Model

As outlined in Sec. 1, we consider the problem of estimating an unknown parameter A from faded observations of sensor encodings generated via quantizer bias control. Specifically, the ith sensor measurement sequence is given by

$$s_i[n] \triangleq A + v_i[n] \ , \tag{1}$$

where the $v_i[n]$'s are independent zero-mean IID Gaussian random processes with power σ_v^2, and A is an unknown parameter that is amplitude-limited in $(-\Delta, \Delta)$. The ith sensor produces a sequential binary-valued encoding $y_i[n]$

$$y_i[n] = \operatorname{sgn}\left(s_i[n] + d_i[n]\right) , \tag{2}$$

where $d_i[n]$ is a control input and $\operatorname{sgn}(\cdot)$ denotes the signum function. We assume that the encodings are communicated over non-interfering time-nonselective frequency-nonselective Rayleigh-fading channels, so that after front-end filtering and sampling, the host has available

$$z_i[n] = \alpha_i \sqrt{\mathcal{E}_s}\, y_i[n] + w_i[n] \qquad \text{for } \ i = 1,\, 2,\, \cdots ,\, L \ , \tag{3}$$

where $\mathcal{E}_s$ is the sensor transmit symbol power, the $w_i[n]$'s are independent zero-mean IID circularly symmetric Gaussian random processes with power σ_w^2, and the α_i's are independent zero-mean circularly symmetric Gaussian random variables each with power σ_α^2. Also, for any i, j, k the processes $w_i[n]$, $v_j[n]$ and the α_k's are assumed to be statistically independent. It is also assumed the α_i's are known to the host but not the sensors.

The fusion task at the host can be summarized as follows: estimate A from the $z_i[n]$'s for $n = 1, 2, \cdots, N$, and $i = 1, 2, \cdots, L$, given the available host information about the control input sequences used at the sensors, and knowledge of the fading coefficients α_i. The sensor objective is to optimize the associated control input strategy so as to minimize the resulting MSE at the fusion center.

2.1 Encoding and Fusion Performance Metrics

As in [1, 5], the quality of the sensor encoding is evaluated via the information loss induced by the encoding and the channel, that is, via the ratio of the Cramér-Rao bounds for unbiased estimates of the parameter A obtained via $z_i[n]$ and $s_i[n]$, respectively, *i.e.*,

$$\mathcal{L}(A; \mathbf{Z}^N) \triangleq \frac{\mathcal{B}\left(A;\ \mathbf{Z}^N\right)}{\mathcal{B}\left(A;\ \mathbf{S}^N\right)}, \tag{4}$$

where $\mathcal{B}\left(A;\ \mathbf{Z}^N\right)$ is the Cramér-Rao bound[1] [6] for unbiased estimation of A from $\mathbf{Z}^N$,

$$\mathbf{Z}^N \triangleq \left[\left(\mathbf{z}_1^N\right)^T\ \left(\mathbf{z}_2^N\right)^T\ \cdots \left(\mathbf{z}_L^N\right)^T\right]^T, \tag{5a}$$

$$\mathbf{z}_i^N \triangleq \left[\mathbf{z}_i[1]\ \ \mathbf{z}_i[2]\ \ \cdots \mathbf{z}_i[N]\right]^T, \tag{5b}$$

and where $\mathbf{S}^N$ and $\mathcal{B}\left(A;\ \mathbf{S}^N\right)$ are defined similarly. We note that better systems achieve smaller information loss over the range of parameter values of interest. Since $\mathcal{B}\left(A;\ \mathbf{S}^N\right) = \sigma_v^2/NL$, both $\mathcal{B}\left(\Delta;\ \mathbf{Z}^N\right)$ and $\mathcal{L}(\Delta; \mathbf{Z}^N)$ can be used interchangeably as figures of merit for assessing the encoding fidelity.

Given that the signal parameter is assumed to be unknown, worst-case performance is used to characterize the overall system; we define the worst-case information loss via

$$\mathcal{L}_{\max}(\Delta; \mathbf{Z}^N) \triangleq \sup_{|A|<\Delta} \mathcal{L}(A; \mathbf{Z}^N). \tag{6}$$

The worst-case information loss is a function of other system parameters, such as σ_v, $\mathcal{E}_s$, the control inputs, and the fading coefficients $\{\alpha_i\}_{i=1}^L$, the dependence on which is suppressed for convenience in the above definitions.

A useful measure of performance of a specific system comprising an encoder and a given estimator is the *MSE loss*, which is defined as the ratio of the actual MSE of a given estimator of A from $\mathbf{Z}^N$, divided by $\mathcal{B}\left(A;\ \mathbf{S}^N\right)$. Since an efficient estimator of A based on $\mathbf{S}^N$ exists (*e.g.*, the sample mean), the notion of the MSE loss of any given estimator of A given $\mathbf{Z}^N$ has a very useful interpretation: it represents the additional MSE in dB that arises from estimating A using a particular estimator on $\mathbf{Z}^N$, instead of efficiently estimating A via $\mathbf{S}^N$. Analogously to $\mathcal{L}_{\max}$ in (6), the worst-case MSE loss of an estimator is defined as the supremum of the MSE loss function over the range $|A| < \Delta$.

---

[1] The use of the term information loss follows from the fact that (4) also equals the inverse of the ratio of the associated Fisher information quantities.

3 Optimal Fusion for Random Control Inputs

We first consider the problem of estimation of A from the $z_i[n]$'s given by (3), where the $d_i[n]$'s are zero-mean independent IID Gaussian random processes, independent of $w_i[n]$'s and $v_i[n]$'s, and with power σ_d^2 (assumed to be known for estimation). It is convenient to express the sensor encoding $y_i[n]$ in (2) as

$$y_i[n] = \operatorname{sgn}\left(A + \beta_i[n]\right) ,$$

where

$$\beta_i[n] = v_i[n] + d_i[n] \ . \tag{7}$$

We note that the $\beta_i[n]$'s are independent zero-mean IID Gaussian random sequences with power $\sigma_\beta^2 = \sigma_v^2 + \sigma_d^2$.

The maximum likelihood (ML) estimate of A based on $z_i[n]$ for $1 \leq i \leq L$ and $1 \leq k \leq N$ over the restricted dynamic range $|A| \leq \Delta$ satisfies[2]

$$\hat{A}_{\mathrm{ML}}\left(\mathbf{Z}^N; \Delta\right) = \arg\max_{|\theta| \leq \Delta} \ \ln P\left(\mathbf{Z}^N; \theta\right) , \tag{8}$$

where $\ln P\left(\mathbf{Z}^N; \theta\right)$ is the associated log-likelihood function. An Estimate-Maximize (EM) algorithm [7] can be employed to obtain $\hat{A}_{\mathrm{ML}}\left(\mathbf{Z}^N; \Delta\right)$. Specifically, by identifying $\mathbf{Z}^N$ as the incomplete data and $\{\mathbf{Z}^N, \mathbf{Y}^N\}$ as the complete data (where $\mathbf{Y}^N$ is defined similarly to $\mathbf{Z}^N$), we obtain the following recursion for the EM algorithm estimates

$$\hat{A}_{\mathrm{EM}}^{(k+1)} = -\mathcal{I}_\Delta\left(\sigma_\beta Q^{-1}\left(\frac{1}{L\,N}\sum_{i=1}^{L}\sum_{n=1}^{N}\frac{g(z_i[n] - \alpha_i\sqrt{\mathcal{E}_{\mathrm{s}}}; \sigma_w)\, Q\left(-\frac{\hat{A}_{\mathrm{EM}}^{(k)}}{\sigma_\beta}\right)}{\sum_{j=\pm 1} g(z_i[n] - j\alpha_i\sqrt{\mathcal{E}_{\mathrm{s}}}; \sigma_w)\, Q\left(-j\frac{\hat{A}_{\mathrm{EM}}^{(k)}}{\sigma_\beta}\right)}\right)\right) , \tag{9}$$

where $Q(x) = \int_x^\infty (1/\sqrt{2\pi})\, e^{-t^2/2}\, dt$, $Q^{-1}(\cdot)$ is the inverse of $Q(\cdot)$, $g(x, \sigma)$ denotes the PDF of a circularly symmetric zero-mean Gaussian random variable with power σ^2, *i.e.*,

$$g(x, \sigma) \triangleq \frac{1}{2\pi\sigma^2}\, e^{-|x|^2/\sigma^2} ,$$

and $\mathcal{I}_\Delta(\cdot)$ is the following piecewise-linear limiter function

$$\mathcal{I}_\Delta(x) = \begin{cases} x & \text{if } \ |x| \leq \Delta \\ \Delta \ \operatorname{sgn}(x) & \text{otherwise} \end{cases} . \tag{10}$$

There are a number of limitations with this approach. First, the derivation of this EM algorithm relies on the existence of a *closed form* solution for the ML estimate of A given the *original* sensor encodings $y_i[n]$. No closed form solution for the ML estimate of A based on the $y_i[n]$'s exists for known time-varying control inputs (such as those employed in feedback control), or M-ary valued quantizer-bias control encoders with $M > 2$. In addition, unlike the cases examined in [1, 5], the worst-case information loss (6) is a computationally intractable quantity to optimize over σ_d. For these reasons we next turn our attention to two different types of approaches which attempt to address the two aforementioned problems; that of an information loss metric that would facilitate the design of the encoder, and that of an algorithmic framework for fusion.

---

[2] Estimators of A for the control-free case can be readily obtained as a special case of the random control input case by setting $\sigma_d = 0$.

4 Encoding and Fusion with Detector Front-End

In this section we examine fusion methods that are constrained to use a front-end detector for each sensor encoding prior to fusion. As the host knows the α_i's, it may employ optimal detection of $y_i[n]$ given $z_i[n]$, *i.e.*, obtain

$$\hat{y}_i[n] = \operatorname{sgn}\left(\operatorname{Re}\{z_i[n]\,\alpha_i^*[n]\}\right) .$$

In terms of the fusion rule, the system with input $y_i[n]$ and output $\hat{y}_i[n]$ is equivalent to a binary symmetric channel with error probability

$$P(\alpha_i) \triangleq \Pr(\hat{y}_i[n] \neq y_i[n]) = Q\left(\sqrt{2\xi}|\alpha_i|\right) , \tag{11}$$

where $\xi = \mathcal{E}_s/\sigma_w^2$. Note that the $P(\alpha_i)$'s are available to the host for fusion.

4.1 Random Control Inputs

In this section we consider random control inputs, *i.e.*, the case where the $d_i[n]$'s correspond to sample paths of independent zero-mean IID Gaussian random processes each with power σ_d^2 that are independent of the $v_i[n]$'s, α_i's, and $w_i[n]$'s. The objective in encoder design is to select the control input power σ_d^2 so as to minimize the worst-case information loss $\mathcal{L}_{\max}(\Delta, \hat{\mathbf{Y}}^N)$ in the detected encodings $\hat{\mathbf{Y}}^N$, where $\hat{\mathbf{Y}}^N$ is defined similarly to $\mathbf{Z}^N$ in (5), and $\mathcal{L}_{\max}(\Delta, \hat{\mathbf{Y}}^N)$ is given by (6) with $\mathbf{Z}^N$ replaced by $\hat{\mathbf{Y}}^N$.

Since the sequences $w_i[n]$ and $\beta_i[n]'s$ in (7) are independent IID sequences and independent of the α_i's, the information loss $\mathcal{L}(A; \hat{\mathbf{Y}}^N)$ satisfies

$$\mathcal{L}(A; \hat{\mathbf{Y}}^N) = L\left[\sum_{i=1}^{L} [\mathcal{L}(A;\hat{y}_i)]^{-1}\right]^{-1} , \tag{12a}$$

where $\mathcal{L}(A;\hat{y}_i)$ corresponds to the information loss in estimating A based on any one sample of the IID sequence $\hat{y}_i[n]$ instead of $s_i[n]$, and is given by [5]

$$\mathcal{L}(A;\hat{y}_i) = \frac{2\,\pi\,\sigma_\beta^2}{(1-2\,P(\alpha_i))^2\sigma_v^2}\,\exp\left(\frac{A^2}{\sigma_\beta^2}\right) C_{\alpha_i}\left(-\frac{A}{\sigma_\beta}\right)\left[1 - C_{\alpha_i}\left(-\frac{A}{\sigma_\beta}\right)\right] , \tag{12b}$$

with $P(\alpha)$ given by (11), and

$$C_\alpha(x) \triangleq (1-2P(\alpha))\,Q(x) + P(\alpha) . \tag{12c}$$

The task of selecting σ_d^2 so as to minimize $\mathcal{L}_{\max}(\Delta; \hat{\mathbf{Y}}^N)$ is at the sensors. From (12), however, $\mathcal{L}_{\max}(\Delta; \hat{\mathbf{Y}}^N)$ and hence $\mathcal{L}_{\max}(\Delta; \hat{\mathbf{Y}}^N)$ depend on the α_i's, which are not available to the sensors. Alternatively, we may consider using $\sup_{|A|<\Delta} E\left[\mathcal{L}(A;\hat{\mathbf{Y}}^N)\right]$ as the performance metric where the expectation is over the PDF of $\{\alpha_i\}_{i=1}^L$. This involves computation of a $2\,L$-fold integral and is a function of L. As we interested in the large L case, we instead employ the following channel-averaged information loss (valid as $L \to \infty$ via a central limit theorem) which is considerably simpler to compute

$$\bar{\mathcal{L}}(A;\hat{y}) \triangleq \left[E_\alpha\left[\mathcal{L}^{-1}(A;\hat{y})\right]\right]^{-1} = \lim_{L\to\infty} \mathcal{L}(A;\hat{\mathbf{Y}}^N) . \tag{13}$$

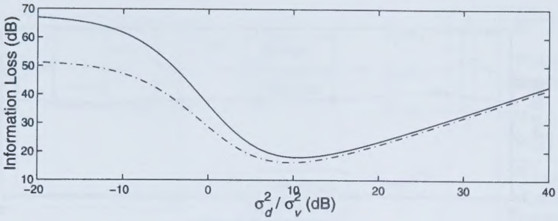


Figure 2: Worst-case information loss a function of the control input power for channel SNRs of 5 dB (solid) and 20 dB (dash-dot); input peak SNR $(\Delta/\sigma_v) = 5$ dB.

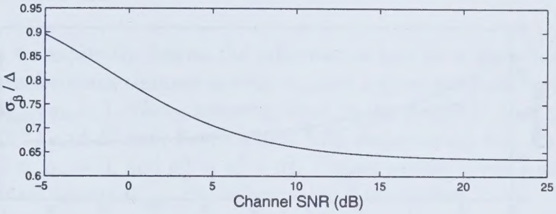


Figure 3: Optimal aggregate noise power level vs. channel SNR.

As the worst-case channel-averaged information loss satisfies

$$\bar{\mathcal{L}}_{\max}(\Delta; \hat{y}) \triangleq \sup_{|A|<\Delta} \bar{\mathcal{L}}(A; \hat{y}) = \bar{\mathcal{L}}(\Delta; \hat{y}) \ , \tag{14}$$

the encoder's objective reduces to selecting σ_d^2 so as to minimize $\bar{\mathcal{L}}(\Delta; \hat{y})$.

Fig. 2 depicts $\bar{\mathcal{L}}_{\max}(\Delta, \hat{y})$ in (14) as a function of the σ_d^2/σ_v^2 and for two different values of the channel SNR $\zeta = \mathcal{E}_s \sigma_\alpha^2 / \sigma_w^2$. The figure reveals that there is an optimal control power level σ_d^2 which depends on the channel SNR. This is further verified by Fig. 3 depicting the optimal aggregate power level σ_β^2 as a function of the channel SNR. The figure also shows that in order to accurately optimize the quality of the random control encoding at the sensors it is indeed important to take into account the quality of the fading channels.

4.1.1 Estimation

The ML estimate of A given $\hat{\mathbf{Y}}^N$ can be obtained by means of an EM algorithm. By identifying $\mathbf{S}^N + \mathbf{D}^N$ (where $\mathbf{D}^N$ is defined similarly to $\mathbf{Z}^N$) as the complete data, the recursion for the resulting EM algorithm is given by

$$\hat{A}_{\mathrm{EM}}^{(k+1)}[N] = \mathcal{I}_\Delta \left(\hat{A}_{\mathrm{EM}}^{(k)}[N] + \frac{1}{LN} \sum_{i=1}^{L} \sum_{n=1}^{N} V(\hat{y}_i[n], \alpha_i, \hat{A}_{\mathrm{EM}}^{(k)}[N]) \right) , \tag{15a}$$

where

$$V(x, \alpha, \theta) = \frac{\sigma_\beta \left(1 - 2\, P(\alpha)\right) x \, \exp\left(-\frac{\theta^2}{2\sigma_\beta^2}\right)}{\sqrt{2\pi}\, C_\alpha \left(-\frac{x\,\theta}{\sigma_\beta}\right)} , \tag{15b}$$

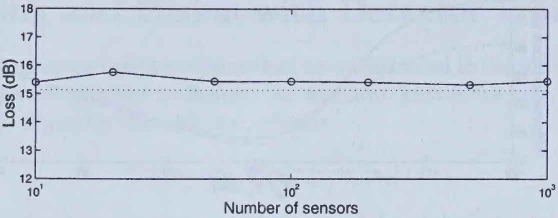


Figure 4: MSE loss ('o') for the algorithm (15), and bound (13) (dot) as a function of the number of sensors; $A/\sigma_v = 4$, $A/\sigma_d = 4/3$, channel SNR=10 dB, $LN = 4000$.

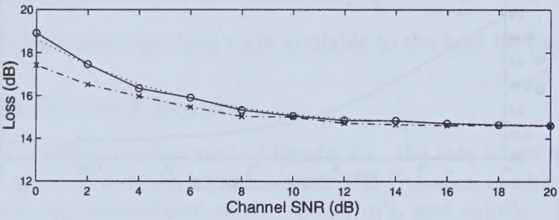


Figure 5: MSE loss ('o') for the algorithm (15), the algorithm (9) ('x'), and bound (13) (dot) as a function of channel SNR; $A/\sigma_v = 4$, $A/\sigma_d = 4/3$, $L = 50$, $N = 80$.

and where $P(\alpha)$ and $C_\alpha(x)$ are given by (11) and (12c), respectively.

Fig. 4 depicts the MSE loss of the algorithm (15) based on Monte-Carlo simulations as a function of L, and $\bar{\mathcal{L}}(A; \hat{y})$ in (13). As we can see the bound (13) obtained by letting $L \to \infty$ accurately predicts the MSE performance of the algorithm (15) even for $L = 10$.

Fig. 5 depicts the bound (13) and the MSE loss of the algorithm (15) based on Monte-Carlo simulations as a function of channel SNR. As the figure reveals, the bound (13) accurately predicts the MSE performance of the algorithm (15) over a wide range of channel SNRs. Also depicted in the figure is the MSE loss of the algorithm (9) operating on the $z_i[n]$'s ('x') based on Monte-Carlo simulations. As the figure reveals, use of a front-end detector incurs only a small performance loss, especially at high SNR.

4.2 Control Inputs in the Presence of Feedback

In this section we consider the feedback control case, where the host broadcasts to the sensors control adjustment information based on past received encodings. The setting is depicted in Fig. 6, whereby the control input used by the ith sensor at time n is given by

$$d_i[n] = d_{\text{host}}[n] + u_i[n] \, ,$$

where the $u_i[n]$'s represent the broadcast channel distortion and are modeled as independent zero-mean IID real-valued Gaussian random processes with power σ_u^2, and $d_{\text{host}}[n]$ represents the host control adjustment sequence computed from past encodings, *i.e.*,

$$d_{\text{host}}[n] = h(\hat{\mathbf{Y}}^{n-1}) \, .$$

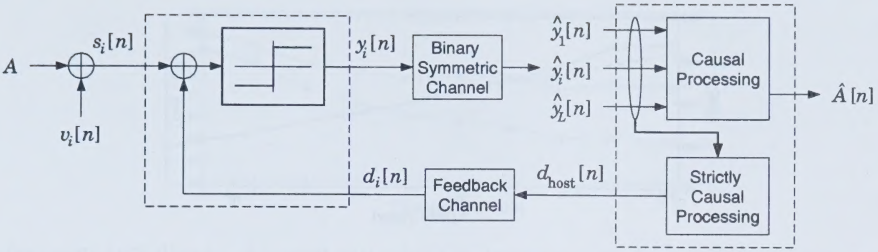


Figure 6: Quantizer bias control with feedback over fading channels and fusion with encoding detector front-end.

In this section we explicitly denote the information loss for a given host control input vector $\mathbf{d}_{\text{host}}^N$, a given feedback channel quality σ_u, and a given random control input power σ_d^2 as $\mathcal{L}(A; \hat{\mathbf{Y}}^N, \mathbf{d}_{\text{host}}^N, \sigma_d, \sigma_u)$. Note, however, that in the feedback case we have $\sigma_d = 0$. We reserve $\mathcal{L}(A; \hat{\mathbf{Y}}^N)$ and $\bar{\mathcal{L}}(A; \hat{y})$ for (12) and (13), respectively, *i.e.*, for the case where $d_{\text{host}}[n] = 0$ for all n, $\sigma_u = 0$, and $\sigma_\beta^2 = \sigma_v^2 + \sigma_d^2$. Consequently, given a set $\{\alpha\}_{i=1}^L$, and a vector of host control inputs $\mathbf{d}_{\text{host}}^N$, the information loss satisfies

$$\begin{aligned}\mathcal{L}(A; \hat{\mathbf{Y}}^N, \mathbf{d}_{\text{host}}^N, 0, \sigma_u) &= \mathcal{L}(A; \hat{\mathbf{Y}}^N, \mathbf{d}_{\text{host}}^N, \sigma_u, 0) = LN \left[\sum_{i,n} \mathcal{L}(A + d_{\text{host}}[n]; \hat{y}_i[n], 0, \sigma_u, 0) \right]^{-1} \\ &\geq \mathcal{L}(0; \hat{\mathbf{Y}}^N) \,. \end{aligned} \tag{16}$$

Ineq. (16) is due to fact that $\mathcal{L}(A; \hat{y}_i[n], 0, \sigma_d, 0)$ is an increasing function of $|A|$; equality holds if and only if $d_{\text{host}}[n] = -A$. This control input results in

$$\bar{\mathcal{L}}(A; \hat{y}, d_{\text{host}}) \geq \bar{\mathcal{L}}(0; \hat{y}) \,, \tag{17}$$

where $\bar{\mathcal{L}}(A; \hat{y}, d_{\text{host}})$ is the worst-case channel-averaged information loss for a given host control input $d_{\text{host}}[n]$ and where $\bar{\mathcal{L}}(0; \hat{y})$ is given by (13) with $\sigma_\beta^2 = \sigma_u^2 + \sigma_v^2$.

4.2.1 Estimation

Although the control input choice $d_{\text{host}}[n] = -A$ that minimizes (17) is not permissible as it depends on A, it suggests viable control input selection methods based on past samples of $\hat{y}_i[n]$. Specifically, if $\hat{A}_{\text{ML}}[n]$, the ML estimate of A based on $\hat{\mathbf{Y}}^n$, is employed for fusion at the host, a reasonable choice for the control input sequence is as follows

$$d_{\text{host}}[n] = -\hat{A}_{\text{ML}}[n-1] \,. \tag{18}$$

The ML estimate $\hat{A}_{\text{ML}}[n]$ can be obtained via the following EM algorithm,

$$\hat{A}_{\text{EM}}^{(k+1)}[n] = \mathcal{I}_\Delta \left(\hat{A}_{\text{EM}}^{(k)}[n] + \frac{1}{Ln} \sum_{i=1}^{L} \sum_{\ell=1}^{n} V(\hat{y}_i[\ell], \alpha_i, \hat{A}_{\text{EM}}^{(k)}[n] + d_{\text{host}}[\ell]) \right) \,, \tag{19}$$

where $V(\cdot\,, \cdot, \cdot)$ is given by (15b), $d_{\text{host}}[\ell]$ is given by (18), and where (19) is initialized with $\hat{A}_{\text{EM}}^{(0)}[n] = \hat{A}_{\text{ML}}[n-1]$ and $\hat{A}_{\text{ML}}[0] = 0$.

Fig. 7 depicts the MSE (solid) of the ML estimator (19) with $d_{\text{host}}[n]$ from (18), based on Monte-Carlo simulations as a function of n, plotted against the bounds $\bar{\mathcal{L}}(0, \hat{y})\sigma_v^2/(nL)$, and $\mathcal{B}(A;\ \mathbf{S}^n) = \sigma_v^2/(nL)$. As the figure reveals, the ML estimate (19) used in conjunction with $d_{\text{host}}[n]$ in (18) effectively achieves the bound (17) provided n is large enough.

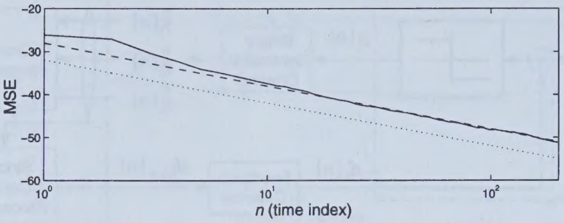


Figure 7: $\bar{\mathcal{L}}(0;\hat{y})\,\sigma_v^2/(nL)$ (dash), $\mathcal{B}\,(0;\ \mathrm{S}^n)$ (dot), and MSE (solid) from Monte-Carlo simulations for $\hat{A}_{\mathrm{ML}}[n]$ in (19). Parameters: $\zeta = 20$ dB, $L = 40$, $\sigma_v = 0.16$, $\sigma_u = 0.12$.

5 Fusion with Linear Combining Front-End

In this section we consider the case of random control inputs, whereby the fusion front-end consists of linear combining of the $z_i[n]$'s. Specifically, fusion is based on the statistic

$$u = \sum_{i=1}^{L} \lambda_i \left[\frac{1}{N} \sum_{n=1}^{N} z_i[n] \right] , \tag{20a}$$

and where the estimate is given by $\hat{A} = F(u)$. The objective reduces to determining $F(\cdot)$ and the λ_i's so as to minimize the $E\left[(A - \hat{A})^2; A\right]$.

First we consider the choice of $F(\cdot)$. For high channel SNR and large L, we have

$$E\,[u; A] \approx \left(2\,Q\left(-A/\sigma_\beta\right) - 1\right) \sum_{i=1}^{L} \lambda_i\, \alpha_i .$$

We consequently employ the following asymptotically unbiased estimator, which also takes into account the range limitation $|A| < \Delta$,

$$\hat{A} = F(u) = -\mathcal{I}_\Delta \left(\sigma_\beta\, Q^{-1} \left(\mathrm{Re} \left\{ \frac{u}{2 \sum_{i=1}^{L} \lambda_i\, \alpha_i} + \frac{1}{2} \right\} \right) \right) . \tag{20b}$$

At high channel SNR and for large L, $E\left[\hat{A}(u); A\right] \approx A$ and

$$\mathrm{var}\,(\hat{A}; A) \approx \frac{1}{N} \mathcal{B}\,(A;\ y)\, \frac{\sum_{i=1}^{L} |\lambda_i|^2\, \left(|\alpha_i|^2 + \bar{b}(A)\right)}{\left| \sum_{i=1}^{L} \lambda_i\, \alpha_i \right|^2} , \tag{21}$$

where $\bar{b}(A) = \left(8\,Q\left(-A/\sigma_\beta\right)\, Q\left(A/\sigma_\beta\right)\right)^{-1}$ and $\mathcal{B}\,(A;\ y)$ is the Cramér-Rao bound for estimating A from $y_i[n]$. The λ_i's that minimize the right hand side of (21) can be obtained via application of the Schwartz inequality as $\lambda_i \propto G_i(A)$, where

$$G_i(A) = \frac{\alpha_i^*}{|\alpha_i|^2 + \bar{b}(A)} , \tag{22}$$

resulting in

$$E\left[(\hat{A} - A)^2; A\right] \approx \frac{1}{N} \mathcal{B}\,(A;\ y) \left[\sum_{i=1}^{L} \frac{|\alpha_i|^2}{|\alpha_i|^2 + \bar{b}(A)} \right]^{-1} .$$

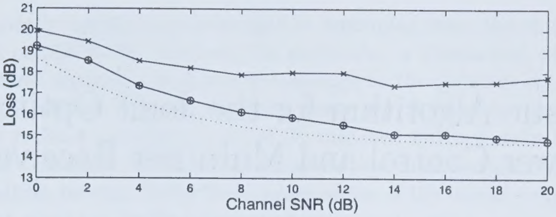


Figure 8: MSE loss for $\hat{A}$ in (20) with one ('x'), two ('o'), and three ('+') iterations, and bound (13) (dot) vs. channel SNR; $A/\Delta = 0.1$, $A/\sigma_v = 4$, $A/\sigma_d = 4/3$, $L = 50$, $N = 80$.

The choice of λ_i's that minimizes $E\left[(\hat{A} - A)^2; A\right]$ depends on knowledge of the unknown parameter A, and hence it is nonvalid. On the other hand, the choice of the λ_i's that minimizes $\sup_{|A|<\Delta} E\left[(\hat{A} - A)^2; A\right]$ is given by $\lambda_i \propto G(\Delta)$ which does not depend on A. Fig. 8 shows the MSE performance of the algorithm (20) with $\lambda_i \propto G_i(\Delta)$, for a particular value of A. As the figure reveals this method has considerably inferior performance than its ML counterparts (9) and (15) examined in the previous sections. However, performance can be greatly improved by additional iterations of the linear combining method. Specifically given an estimate $\hat{A}^{(1)}$ obtained by the algorithm (20) with $\lambda_i \propto G_i(\Delta)$, an improved estimate $\hat{A}^{(2)}$ can be obtained via a second application of (20) using the improved set $\lambda_i \propto G_i(\hat{A}^{(1)})$. The performance of this iterative linear combining method is also shown in Fig. 8 revealing that significant improvements can be attained via a second iteration but the benefits quickly decline with further iterations.

References

[1] H. C. Papadopoulos, G. W. Wornell, and A. V. Oppenheim, "Sequential signal encoding from noisy measurements using quantizers with dynamic bias control," accepted for publication in *IEEE Trans. Inform. Theory.*

[2] T. Berger, *Rate Distortion Theory.* Prentice-Hall Series in Information and System Sciences, EngleWood Cliffs, New Jersey: Prentice-Hall, 1971.

[3] Y. Ephraim and R. M. Gray, "A unified approach for encoding clean and noisy sources by means of waveform and autoregressive model vector quantization," *IEEE Trans. Inform. Theory,* vol. 34, pp. 826–834, July 1988.

[4] S. P. Lipshitz, R. A. Wannamaker, and J. Vanderkooy, "Quantization and dither: a theoretical survey," *J. Audio Eng. Soc.*, vol. 40, pp. 355–375, May 1992.

[5] H. C. Papadopoulos, "Sequential signal encoding and estimation for wireless sensor networks," in *Proc. Int. Symp. Inform. Theory,* June 2000.

[6] H. L. Van Trees, *Detection, Estimation and Modulation Theory, Part I.* New York, NY: John Wiley and Sons, 1968.

[7] A. P. Dempster, N. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Ann. Roy. Statist. Soc.*, vol. 39, pp. 1–38, Dec. 1977.

A Stochastic Algorithm for the Joint Optimization of Power Control and Multiuser Receivers *

Deepak Das, and Mahesh K. Varanasi
Dept. of Electrical and Computer Engineering
University of Colorado
Boulder, CO 80303
dasd@ucsu.colorado.edu varanasi@schof.colorado.edu

Abstract

Uplink communication in a cellular radio network is considered where the base station in each cell employs a linear or decision feedback (DF) multiuser receiver. The problem of minimizing the total transmit power over the receiver filter coefficients is considered under the constraint that all the users of the network achieve their signal-to-interference ratio (SIR) objectives. A distributed, stochastic algorithm is obtained that converges to the jointly optimal pair of power allocations and maximum SIR decision feedback receivers.

Index Terms: Power control, cellular networks, stochastic approximation, blind receivers.

1 Introduction and System Model

The problem of interest in this paper is that of minimizing the total transmit power under the constraint that all the users of the network achieve their Quality-of-Service (QoS) objective in terms of signal-to-interference ratio (SIR) on the uplink, i. e., at a base station in a cellular system. The minimization of total power required to achieve the SIR objectives is done over all linear and decision feedback (DF) multiuser receivers (both feedforward and feedback filter coefficients in the DF case) to yield the most power efficient solution when the receivers are constrained to be of the linear or decision feedback type. This problem was considered in [1]. A similar problem restricted to the case of linear multiuser receivers was studied in [2]. In this paper, we follow closely the model and problem formulation given in [1].

It is assumed that the decision feedback at each base station involves only decisions made for in-cell users. With complete knowledge of system parameters, the joint multiuser receiver-power optimization essentially consists of alternating between an optimum power computation step and a maximum SIR (MSIR) linear or decision feedback (nonlinear) receiver computation step. Although decentralized and stochastic versions of this iterative optimization algorithm have been proposed in [1] [2], no convergence analysis was presented. In this paper, a different iterative algorithm is introduced for directly

---

*This work was supported in part by NSF grants CCR-9814996 and ANI-9725778.

computing the jointly optimum powers that is decoupled from the computation of the optimum (MSIR or MSIR-DF) receivers. In particular, a distributed, stochastic version of our power control algorithm is shown to converge to the (jointly) optimum powers.

We consider a cellular network in which there are B base stations with K_j users assigned to base j. Base j employs a set of N_j matched filters matched to orthonormal basis functions of a carefully chosen signal subspace that must include the signals received from the users within its cell. While the transmissions of the in-cell mobiles are assumed to be received at the base station symbol-synchronously, those of out-of-cell users are not. Base j is assumed to have knowledge of the common timing of only in-cell users.

The complex baseband discrete-time model for the N_j matched filter outputs at base j can be expressed as

$$\mathbf{y}_j = \sum_{i=1}^{K_j} \sqrt{w_{ij} g_{ijj}} \mathbf{s}_{ijj} x_{ij} + \sum_{l \neq j}^{B} \sum_{i=1}^{K_l} \sqrt{w_{il} g_{ilj}} \left(\mathbf{s}_{ilj}^{-} x_{il}^{-} + \mathbf{s}_{ilj}^{+} x_{il}^{+} \right) + \mathbf{n}_j, \tag{1}$$

where w_{il} and x_{il} denote the transmit power and the transmitted symbol, respectively, of user i of base l. The channel gain of the i^{th} user of base j to base l is denoted by g_{ilj}. The (possibly complex-valued) vector $\mathbf{s}_{ijj}$ denotes the vector representation (the "signature sequence") of the signal of user i of base j with respect to the orthonormal basis functions employed at base j. The vectors $\mathbf{s}_{ilj}^{-}$ and $\mathbf{s}_{ilj}^{+}$ denote the vector representations of the segments of the signals associated with the two symbols of user i of base l that overlap with the signals of base j. $\mathbf{n}_j$ is an N_j-dimensional zero-mean, circularly symmetric, Gaussian random vector with a covariance matrix equal to $\sigma_j^2 \mathbf{I}$. Define the signal matrix of in-cell users as $\mathbf{S}_{jj} = \left[\mathbf{s}_{1jj}\mathbf{s}_{2jj}\cdots\mathbf{s}_{K_j jj}\right]$. Also define the signal matrices of out-of-cell users $\mathbf{S}_{lj}^{+} = \left[\mathbf{s}_{1lj}^{+}\mathbf{s}_{2lj}^{+}\cdots\mathbf{s}_{K_l lj}^{+}\right]$ and $\mathbf{S}_{lj}^{-} = \left[\mathbf{s}_{1lj}^{-}\mathbf{s}_{2lj}^{-}\cdots\mathbf{s}_{K_l lj}^{-}\right]$, the gain matrices $\mathbf{G}_{lj} = \operatorname{diag}\{g_{1lj}, g_{2lj}, \cdots, g_{K_l lj}\}$, and power matrices $\mathbf{W}_l = \operatorname{diag}\{w_{1l}, w_{2l}, \cdots, w_{K_l l}\}$.

2 Multiuser Receivers

Each base station uses either a linear or a decision feedback receiver to decode the information transmitted by users in its own cell. In the decision feedback case we assume that the users in each cell are numbered according to the order in which they are decoded. The soft-output of the linear or decision feedback receiver for user i of base j can be written as

$$v_{ij} = \mathbf{f}_{ij}^{\dagger} \left(\mathbf{y}_j - \sum_{k=1}^{i-1} \mathbf{b}_{kij} \sqrt{w_{kj} g_{kjj}} \hat{x}_{kj} \right). \tag{2}$$

"$\dagger$" denotes conjugate-transpose and $\mathbf{f}_{ij}$ is the feedforward filter. The feedback filters $\mathbf{b}_{kij} = \mathbf{0}$ for linear receivers (no decision feedback) and for decision feedback receivers, we let $\mathbf{b}_{kij} = \mathbf{s}_{kjj}$ for $k = 1, \cdots, i-1$, without loss of generality [3]. $\hat{x}_{kj}$ denotes the detected (in uncoded transmission) or decoded (in the coded case) symbols of "past" users $1, \cdots, i-1$ relative to user i of base j. Implicit in (2) is the assumption that base j has knowledge of the codes, signals, transmit powers and channel gains of the users within its own cell. We do not however assume that it has knowledge of these parameters for out-of-cell users. Single-cell receivers, i. e., receivers that require only in-cell information, were studied in [1] as were the MSIR linear and decision feedback receivers.

In this paper we focus on receivers designed for the SIR criterion. First define matrices $\{\mathbf{Q}_{lj}\}$ for $l, j \in \{1, 2, \cdots, B\}$, with $Q_{lj}(i,k)$ denoting the $(i,k)^{th}$ element of $\mathbf{Q}_{lj}$, as follows:

the matrices $\mathbf{Q}_{jj}$ are defined for $j = 1, \cdots, B$ as

$$Q_{jj}(i,k) = \begin{cases} |\mathbf{f}_{ij}^{\dagger}(\mathbf{s}_{kjj} - \mathbf{b}_{kij})|^2 & \text{if } k < i \\ |\mathbf{f}_{ij}^{\dagger}\mathbf{s}_{kjj}|^2 & \text{if } k \geq i \end{cases} \tag{3}$$

for $i, k \in \{1, \cdots, K_j\}$ and the matrices $\mathbf{Q}_{lj}$ for $j, l \in \{1, 2, \cdots, B\}$ with $j \neq l$ are defined as $Q_{lj}(i,k) = |\mathbf{f}_{ij}^{\dagger}\mathbf{s}_{klj}^{-}|^2 + |\mathbf{f}_{ij}^{\dagger}\mathbf{s}_{klj}^{+}|^2$ for $i \in \{1, \cdots, K_j\}$ and $k \in \{1, \cdots, K_l\}$.

Assuming perfect feedback in the case of decision feedback receivers, the SIR in the soft output v_{ij} of user i in base j can be expressed as

$$\gamma_{ij} = \frac{w_{ij} g_{ijj} |\mathbf{f}_{ij}^{\dagger}\mathbf{s}_{ijj}|^2}{\sum_{l=1}^{B} I_{ijl} - w_{ij} g_{ijj} |\mathbf{f}_{ij}^{\dagger}\mathbf{s}_{ijj}|^2 + \sigma_j^2 \mathbf{f}_{ij}^{\dagger}\mathbf{f}_{ij}}\,, \tag{4}$$

where $I_{ijl} = \sum_{k=1}^{K_l} Q_{lj}(i,k) w_{kl} g_{klj}$.

The MSIR receivers, with and without decision feedback (that are also known, with suitable scaling, to minimize mean-squared error) are given by the vectors $\mathbf{f}_{ij}^{msir}$ that satisfy the equations (cf. [4], [3]))

$$\mathbf{A}_{ij}\mathbf{f}_{ij}^{msir} = \mathbf{s}_{ijj} \qquad i = 1, \cdots K_j, \quad j = 1, \cdots, B\,, \tag{5}$$

with

$$\begin{aligned} \mathbf{A}_{ij} \;=\; & \left(\chi \sum_{k=1}^{i-1} g_{kjj} w_{kj} \mathbf{s}_{kjj} \mathbf{s}_{kjj}^{\dagger} + \sum_{k=i}^{K_j} g_{kjj} w_{kj} \mathbf{s}_{kjj} \mathbf{s}_{kjj}^{\dagger} \right. \\ & \left. + \sum_{l \neq j}^{B} \mathbf{S}_{lj}^{-} \mathbf{G}_{lj} \mathbf{W}_l (\mathbf{S}_{lj}^{-})^{\dagger} + \sum_{l \neq j}^{B} \mathbf{S}_{lj}^{+} \mathbf{G}_{lj} \mathbf{W}_l (\mathbf{S}_{lj}^{+})^{\dagger} + \sigma_j^2 \mathbf{I} \right)\,, \end{aligned} \tag{6}$$

where the indicator $\chi = 1$ for linear MSIR and $\chi = 0$ for MSIR-DF. The MSIR-DF receiver can have a markedly improved performance relative to the MSIR linear receiver in both the single- and multi-cell channels [3]. Implicit in this result is the assumption that feedback is perfect in the case of the DF receiver. As has been pointed out in [1], while decision feedback is not perfect in general, it is possible, by decoding the users in the decreasing order of their target SIR's, to mitigate the error propagation effects to a large extent [5]. In the case of equal target SIR's, it may be possible to change the problem slightly and set the target SIR's to be slightly different from each other while ensuring that the worst-case target SIR is equal to the required target SIR. The slight increase in the powers required to achieve these higher set of target SIR's may be a small price to offset the effects of error propagation and achieve better overall performance.

Note that in contrast to the single-cell receivers, the MSIR linear and DF receivers depend on the system parameters corresponding to all users including those outside the cell of interest. Since such parameters are not available at each base station, it is of interest to consider blind adaptive multiuser receivers for the cellular channel (cf. [1]).

3 Joint Receiver-Power Control Optimization

We now consider our central optimization problem - minimizing the minimum total power required to achieve the SIR requirements over all possible linear or decision feedback multiuser receivers.

3.1 Algorithms for Joint Optimization

The joint optimization of multiuser receivers and power control can be formulated (equivalently to the optimization in [1]) as:

$$\min_{\substack{\mathbf{f}_{ij} \in C^N \\ i=1,\cdots,K_j \\ j=1,\cdots,B}} \quad \min_{\substack{\mathbf{w} \geq \mathbf{0} \\ \gamma_{ij}/\gamma_{ij}+1 \geq \gamma^*_{ij}/\gamma^*_{ij}+1}} \sum_{j=1}^{B} \sum_{i=1}^{K_j} w_{ij}, \tag{7}$$

where $\mathbf{w} = [w_{11}, \cdots, w_{K_1 1}, \cdots, w_{1B}, \cdots, w_{K_B B}]^T$ (note that the constraints $\gamma_{ij} \geq \gamma^*_{ij}$ are equivalent to $\frac{\gamma_{ij}}{\gamma_{ij}+1} \geq \frac{\gamma^*_{ij}}{\gamma^*_{ij}+1}$). The above problem can be solved in a synchronous (each base station updates powers and receivers for all its users at the same time), distributed manner by the following deterministic iteration (cf. [1]):

$$\begin{aligned} w_{ij}(n+1) &= \frac{\gamma^*_{ij}\mathbf{f}^\dagger_{ij}(n)\mathbf{A}_{ij}(n)\mathbf{f}_{ij}(n)}{(\gamma^*_{ij}+1)\mathbf{f}^\dagger_{ij}(n)\mathbf{s}_{ijj}\mathbf{s}^\dagger_{ijj}\mathbf{f}_{ij}(n)g_{ijj}}, && (8) \\ \mathbf{f}_{ij}(n+1) &= (\mathbf{A}_{ij})^{-1}(n)\mathbf{s}_{ijj}\,, && (9) \end{aligned}$$

Assuming the target SIRs are feasible (i. e., there exists some feasible power vector for which the SIR constraints can be met), it can be shown that the two-step distributed, synchronous, iterative algorithm (8, 9) converges to the *unique* optimum pair of receivers and power allocations for all users in the network. The proof of convergence for the linear receiver case was shown in [2], and that for the decision feedback receiver case is similar. We will consider here a different algorithm to compute the optimum powers and receivers that will be shown in the following section to be more suitable for a stochastic implementation.

We begin by noting that, given there exists some feasible power vector, the unique solution to the deterministic iteration in (8, 9) is a function only of the user codes, path gains and noise power. In particular, since for user powers $\mathbf{w}$, the SIR achieved with the MSIR receiver is given as $\mathbf{s}^\dagger_{ijj}(\mathbf{A}_{ij})^{-1}\mathbf{s}_{ij}g_{ijj}w_{ij}$ for user i in cell j, the optimum powers $\mathbf{w}^*$, are such that each w^*_{ij} is given by the solution of the nonlinear equation

$$\mathbf{s}^\dagger_{ijj}(\mathbf{A}_{ij})^{-1}\mathbf{s}_{ijj}g_{ijj}w_{ij} = \frac{\gamma^*_{ij}}{1+\gamma^*_{ij}}\,, \text{ and} \tag{10}$$

and the corresponding SIR maximizing MSIR receiver for the $(i,j)^{th}$ user is given in terms of the optimum powers by

$$\mathbf{f}_{ij}{}^* = (\mathbf{A}_{ij}{}^*)^{-1}\mathbf{s}_{ijj}\,. \tag{11}$$

With complete knowledge of system parameters, we can also devise the following deterministic power-receiver recursions for the $(i,j)^{th}$ user

$$\begin{aligned} w_{ij}(n+1) &= w_{ij}(n) - a_n\left[\mathbf{s}^\dagger_{ij}(\mathbf{A}_{ij})^{-1}(n)\mathbf{s}_{ijj}w_{ij}(n)g_{ijj} - \frac{\gamma^*_{ij}}{1+\gamma^*_{ij}}\right], && (12) \\ \mathbf{f}_{ij}(n) &= (\mathbf{A}_{ij})^{-1}(n)\mathbf{s}_{ijj}\,. && (13) \end{aligned}$$

The above two-step joint optimization iterative procedure will be used in the next section to derive the corresponding distributed and stochastic algorithms.

4 Stochastic Joint Optimization

It was mentioned in [2] in the context of linear receivers that one could try to achieve the jointly optimum receiver-power solution by alternately running a blind MSIR receiver adaptation (with the powers fixed) until convergence is achieved, and a stochastic power adaptation (with the receivers fixed) until convergence is achieved. However, switching back and forth between a receiver adaptation and a power adaptation entails waiting at each step for convergence. This may not be feasible within a time frame where the path gains (and hence the optimum joint solution) remain the same. In a practical implementation, where convergence cannot be truly realized at any stage, analytical results that give appropriate stopping criteria are needed so as to guarantee the convergence of this procedure. Unfortunately, no such results are available at this time.

As another attempt at a blind, distributed stochastic joint optimization, the work in [1] suggests a two step stochastic approximation based strategy, one for the power computation and one for the filter computation, in each iteration, to mimic the deterministic iterative solution of (8, 9). Numerical simulations suggest that the combination of the two such coupled adaptations does not necessarily converge to the jointly optimal solution of (8, 9). Therefore, there is a need to analyse the pair of iterations in a framework that provides conditions, say, on the step sizes, under which there is guaranteed convergence in some sense. There does not appear to be an easy way to acomplish this from what is known in the stochastic approximation literature.

We therefore seek a stochastic algorithm that is based on (12, 13) which allows us to compute the jointly optimum powers of the joint receiver-power control optimization solution directly in a manner that is decoupled from the receiver computation. Such an algorithm must be distributed, in the sense that it must be implementable at each base station with a knowledge of only the in-cell parameters available at that base station. It must also be blind in that there should be no need for training sequences.

We form M-length blocks of the received vectors at each base. Let $\mathbf{y}_j(n,m)$ denote the m^{th} vector in the n^{th} block at base j given as

$$\begin{aligned} \mathbf{y}_j(n,m) &= \sum_{i=1}^{K_j} \sqrt{w_{ij}(n) g_{ijj}} \mathbf{s}_{ijj} x_{ij}(n,m) + \mathbf{n}_j(n,m) \\ &+ \sum_{l \neq j}^{B} \sum_{i=1}^{K_l} \sqrt{w_{il}(n) g_{ilj}} \left(\mathbf{s}_{ilj}^{-} x_{il}^{-}(n,m) + \mathbf{s}_{ilj}^{+} x_{il}^{+}(n,m) \right), \end{aligned} \tag{14}$$

where $w_{il}(n)$ denotes the transmit power of user i of base l during the n^{th} block, and $x(n,m)$ with the appropriate sub- and superscripts denote the corresponding transmitted symbols in the m^{th} duration in the n^{th} block. From the M-length block of data, base j computes an unbiased estimate of the terms $\{\mathbf{s}_{ijj}^{\dagger}(\mathbf{A}_{ij})^{-1}\mathbf{s}_{ijj}\}$, $i \in \{1, \cdots, K_j\}$. Let us denote these estimates as $\{\mathbf{s}_{ijj}^{\dagger}(\tilde{\mathbf{A}}_{ij})^{-1}(n)\mathbf{s}_{ijj}\}$, respectively. We now use these estimates in the following power recursion:

$$w_{ij}(n+1) = w_{ij}(n) - a_n \left[\mathbf{s}_{ijj}^{\dagger}(\tilde{\mathbf{A}}_{ij})^{-1}(n)\mathbf{s}_{ijj} w_{ij}(n) g_{ijj} - \frac{\gamma_{ij}^{*}}{1+\gamma_{ij}^{*}} \right]. \tag{15}$$

The step size sequences a_n can be chosen to be fixed or decreasing. The estimates $\{\mathbf{s}_{ijj}^{\dagger}(\tilde{\mathbf{A}}_{ij})^{-1}(n)\mathbf{s}_{ijj}\}$ can be formed in the following way: First we estimate $(\mathbf{A}_{ij})^{-1}$ using the recursion (over $M-1$ iterations within the M-length block):

$$(\hat{\mathbf{A}}_{ij})^{-1}(m+1,n) = (\hat{\mathbf{A}}_{ij})^{-1}(m,n) - \frac{(\hat{\mathbf{A}}_{ij})^{-1}(m,n)\tilde{\mathbf{y}}_{ij}(n,m)\tilde{\mathbf{y}}_{ij}(n,m)^{\dagger}(\hat{\mathbf{A}}_{ij})^{-1}(m,n)}{1+\tilde{\mathbf{y}}_{ij}(n,m)^{\dagger}(\hat{\mathbf{A}}_{ij})^{-1}(m,n)\tilde{\mathbf{y}}_{ij}(n,m)}, \tag{16}$$

where $\tilde{\mathbf{y}}_{ij}(n,m) = \mathbf{y}_j(n,m) - \sum_{k=1}^{i-1} \mathbf{b}_{kij}\sqrt{w_{kj}(n)g_{kjj}}\hat{x}_{kj}(n,m)$. Note that the averaging is done over $M-1$ symbol intervals (ignoring the M^{th} inteval), because the received data in the M^{th} interval reflects, due to asynchronicity of out-of-cell users, two different powers for each out-of-cell user. If M is large, leaving out one symbol interval does not cause any significant loss of efficiency. Moreover, it makes the convergence analysis somewhat simpler. With the final output of the $(M-1)$-length recursion in (16) $(\hat{\mathbf{A}}_{ij})^{-1}(M-1,n)$, we then calculate $\mathbf{s}_{ijj}^{\dagger}(\tilde{\mathbf{A}}_{ij})^{-1}(n)\mathbf{s}_{ijj} = \mathbf{s}_{ijj}^{\dagger}(\hat{\mathbf{A}}_{ij})^{-1}(M-1,n)\mathbf{s}_{ijj}\mu_{bias}(n)$, where $\mu_{bias}(n)$ is chosen to make the estimate unbiased. With the assumption that the received signal is zero-mean Gaussian (as is usually done for the analysis of the Recursive Least Squares (RLS) algorithm) $\mu_{bias}(n)$ is the constant term $\frac{(M-1)-N-1}{M}$.

The MSIR filters corresponding to the optimum powers can be adaptively obtained either concurrently with the power updation, or only when the powers have sufficiently converged. In the latter case, with the powers fixed at $\mathbf{w}(n_N)$, (where $n_N > 0$ is large enough to ensure that the user powers are sufficiently close to the optimum powers), the MSIR filters can be obtained blindly using well known adaptive algorithms (cf. [1]). For blindly estimating the optimum MSIR receivers $\mathbf{f}_{ij}^{*}$ concurrently with the power updation, one can use the following stochastic approximation based recursion indexed by (n,m), where m is the updation number within the n^{th} $(M-1)$-length block of received vectors (recall that the powers are updated only after every $(M-1)$-length interval and, therefore, remain constant for each block):

$$\mathbf{f}_{ij}(n,m+1) = \mathbf{f}_{ij}(n,m) - b_n\left[\tilde{\mathbf{y}}_{ij}(n,m)\tilde{\mathbf{y}}_{ij}(n,m)^{\dagger}\mathbf{f}_{ij}(n,m) - \mathbf{s}_{ijj}\right] . \tag{17}$$

The step size b_n is chosen to be a sufficiently small fixed number, such that the filter recursion in (17) can track the changing powers from (15) and estimate the optimum filters within an acceptable asymptotic mean squared error.

5 Convergence Analysis

We shall first show that the deterministic iteration in (12) converges to $\mathbf{w}^{*}$, the optimum powers of the joint receiver-power control optimization. Let the vector valued function

$$\begin{aligned} \mathbf{p}(\mathbf{w}(n)) &\triangleq diag\left[\mathbf{s}_{111}^{\dagger}(\mathbf{A}_{11})^{-1}(n)\mathbf{s}_{111}g_{111}, \cdots, \mathbf{s}_{K_BBB}^{\dagger}(\mathbf{A}_{K_BB})^{-1}(n)\mathbf{s}_{K_BBB}g_{K_BBB}\right]\mathbf{w}(n) , \\ &\triangleq \Theta(n)\mathbf{w}(n) , \end{aligned} \tag{18}$$

and $\Gamma \triangleq \left[\frac{\gamma_{11}^{*}}{1+\gamma_{11}^{*}}, \cdots, \frac{\gamma_{K_BB}^{*}}{1+\gamma_{K_BB}^{*}}\right]^T$. Clearly, $\Theta(n)$ is a function of $\mathbf{w}(n)$. If we denote its value for $\mathbf{w}(n) = \mathbf{w}^{*}$ as Θ^{*}, then $\Theta^{*}\mathbf{w}^{*} = \Gamma$. For $\mathbf{w} > 0$, we note that the function $\mathbf{p}(\mathbf{w})$ is continuously differentiable and denote its derivative by $\mathbf{p}'(\mathbf{w})$. Let us consider the case when the step-size $a_n = a > 0$.

Theorem 1 *When the joint receiver-power control optimization has a solution, there exists $\bar{a} > 0$ such that when $0 < a < \bar{a}$, the algorithm in (12) is locally convergent to $\mathbf{w}^{*}$.*

Proof: Let us define $\epsilon(n) = \mathbf{w}(n) - \mathbf{w}^*$ and let $||\epsilon(n)||$ be its norm. We can locally expand $\mathbf{p}(\mathbf{w}(n))$ as

$$\mathbf{p}(\mathbf{w}(n)) = \mathbf{p}(\mathbf{w}^*) + \mathbf{p}^{'}(\mathbf{w}^*)\epsilon(n) + o(||\epsilon(n)||) \ , \tag{19}$$

and rewrite the recursion in (12) as

$$\epsilon(n+1) \ = \ \mathbf{D}\epsilon(n) + o(||\epsilon(n)||) \ , \tag{20}$$

where $\mathbf{D} \triangleq \mathbf{I} - a\mathbf{p}^{'}(\mathbf{w}^*)$. We can show that $\mathbf{p}^{'}(\mathbf{w}^*)$ is an M-matrix [8] and specifically, it is *positive stable*, i. e., the real part of each eigenvalue of $\mathbf{p}^{'}(\mathbf{w}^*)$, Re $\lambda_k(\mathbf{p}^{'}(\mathbf{w}^*))$, is positive. Now consider $\bar{a} = 2 \ \min\left[\text{Re } \lambda(\mathbf{p}^{'}(\mathbf{w}^*))/|\lambda(\mathbf{p}^{'}(\mathbf{w}^*))|^2\right]$, which, by the positive-stableness of $\mathbf{p}^{'}(\mathbf{w}^*)$, is positive too. Let $\mu_k(\mathbf{D})$ be the k^{th} eigenvalue of $\mathbf{D}$, so that $\mu_k(\mathbf{D}) = 1 - a\lambda_k(\mathbf{p}^{'}(\mathbf{w}^*))$. Therefore, when $0 < a < \bar{a}$, and denoting $\mu_k(\mathbf{D})$ simply as μ_k and $\lambda_k(\mathbf{p}^{'}(\mathbf{w}^*))$ as λ_k, we have

$$|\mu_k|^2 = |1 - a\lambda_k|^2 = 1 + a^2|\lambda_k|^2 - 2a\text{Re}\lambda_k = 1 - a|\lambda_k|^2\left(\frac{2\text{Re } \lambda_k}{|\lambda_k|^2} - a\right) < 1. \tag{21}$$

But, if all the eigenvalues of $\mathbf{D}$ have modulus less than 1, we have $||\epsilon(n)|| \rightarrow 0$ at the rate of a geometric progression when $||\epsilon(0)||$ is small. ◇

We shall now address the important problem of convergence of the stochastic power control algorithm in (15). We first make the assumption that our estimates $\{\mathbf{s}_{ijj}^{\dagger}(\tilde{\mathbf{A}}_{ij})^{-1}(n)\mathbf{s}_{ijj}\}$ are unbiased. Similar to the definition of $\Theta(n)$ in (18), we define the matrix $\tilde{\Theta}(n) \ = \ diag\left[\mathbf{s}_{111}^{\dagger}(\tilde{\mathbf{A}}_{11})^{-1}(n)\mathbf{s}_{111}g_{111}, \cdots, \mathbf{s}_{K_BBB}^{\dagger}(\tilde{\mathbf{A}}_{K_BB})^{-1}(n)\mathbf{s}_{K_BBB}g_{K_BBB}\right]$ and the vector $\xi(n) \ \triangleq \ \tilde{\Theta}(n)\mathbf{w}(n) \ - \ \Theta(n)\mathbf{w}(n)$. Let the error vector be defined as $\epsilon(n) \ = \ \mathbf{w}(n) - \mathbf{w}^*$ and the vector $\zeta(n) \ \triangleq \ \Theta(n)\mathbf{w}(n) - \Gamma$. The recursion in (15) can be rewritten in terms of the error, as

$$\epsilon(n+1) \ = \ \epsilon(n) - a_n\left[\zeta(n) + \xi(n)\right] . \tag{22}$$

We consider the proof technique in [7] to show *local* convergence for our algorithm with the following standard conditions on the step-size sequence: $\sum_n a_n = \infty$, $\sum_n a_n^2 < \infty$ (note that the technique in [9] does not seem applicable). We next define the set $\Omega = \{\epsilon(n) : ||\epsilon(n)||^2 \leq v\}$ for all n, where $v > 0$ is some number such that a local expansion of the vector function $\Theta(n)\mathbf{w}(n)$ in (19) is valid. We then use Lyapunov's theorem for stable matrices to get suitable bounds on quadratic terms involving $\epsilon(n)$ in order to satisfy the conditions in [7] that make the following result applicable in proving local convergence:

Theorem 2 *When the joint receiver-power control optimization has a solution, then the following is true of the algorithm in (15):*

$$P(\epsilon(n) \in \Omega \ \forall \ n) \ \geq \ 1 - \delta \ , \ \delta = v^{-1}\left(||\epsilon(0)||^2 + K\sum_{i=0}^{\infty} a_i^2\right) \ , \tag{23}$$

$$\tag{24}$$

where $0 < K < \infty$.

From the local convergence result in Theorem 2, we can deduce that δ is small, when the domain in which the conditions of the theorem are fulfilled, is large, i. e., v is large, when the initial approximation is good, i. e., $||\epsilon(0)||^2$ is small, and when the step size sequence is small, i. e., $\sum a_n^2$ is small.

6 Numerical Examples

Example 1 *We consider the 7-cell cellular system shown in Figure 1 with 20 users in each cell. The noise power is about 20 dB less than the average received signal power. The processing gain is 64, and the signature sequences are unit norm random vectors. Once chosen, the signature sequences, remain fixed throughout the adaptation. The target SIR for each user is 8. For this example, the (deterministic) jointly optimum sum of powers for the linear MSIR based power control is a little more than 1.2 times the power required by the MSIR-DF based power control (a saving of 20% with DF).*

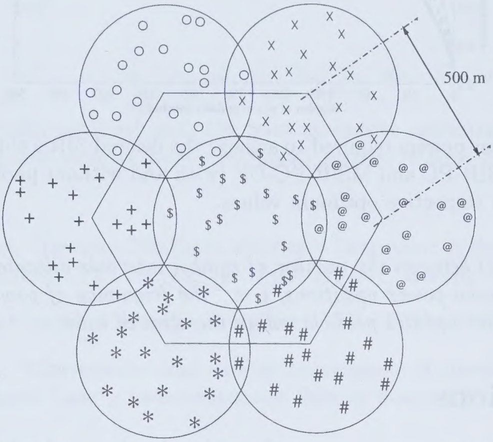


Figure 1: 7-cell cellular system with 20 users per cell

The powers are updated every 100-length block of symbol intervals for both the MSIR-DF-PC (with and without perfect feedback) and MSIR-PC algorithms. The filters are estimated concurrently with the powers as described in (17).

Figure 2 shows the estimated sum of user powers for both the linear MSIR and MSIR-DF based power control algorithms relative to their corresponding optimum (deterministic) values plotted against the iteration index.

Figure 3 plots the average (over users) SIR performance of the estimated MSIR and MSIR-DF receivers (relative to the desired common SIR of 8). Both the power control and the receiver algorithms were run with suitably chosen fixed step sizes, the step sizes being the same for the linear MSIR and MSIR-DF cases.

It is clear from the figure that the estimated sum of powers as well as the average SIR converge to their respective desired values. Note that although the MSIR-DF approach with perfect feedback assumption is more efficient than the linear MSIR approach (both in terms of sum of powers required and convergence rate of the stochastic algorithms), the effects of error propagation in the DF case does have a slightly adverse effect. This example with equal SIR's is, however, the worst case scenario in terms of error propagation. It was also observed over different numerical examples that when the powers are updated less frequently and estimates of $(\tilde{\mathbf{A}}_{ij}^{DF})^{-1}(n)$ *are computed over more symbol intervals (yielding, therefore, better estimates), the convergence (for a given number of power updates) improves. Since it is desirable to get acceptably close to the optimum powers over a limited number of symbol intervals there is a trade-off (in determining the*

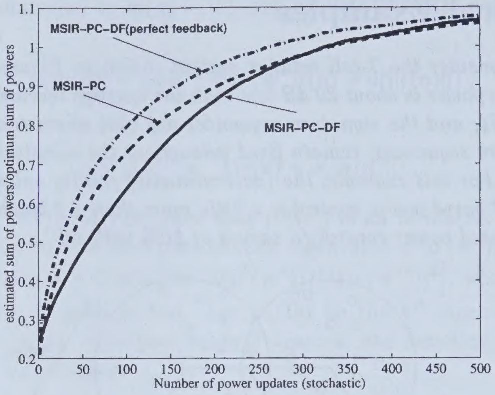


Figure 2: Sum of user powers required to achieve the desired SIR's with the joint stochastic optimization MSIR-PC and MSIR-PC-DF (with and without perfect feedback), plotted *relative* to their respective optimum values.

overall performance) between the number of symbol intervals allocated to the estimation of $(\tilde{\mathbf{A}}_{ij}^{DF})^{-1}(n)$ *between power updations, i. e., the frequency of power updates, and the total number of power updates possible within the allowed limit.*

7 Conclusions

The joint multiuser receiver-power control problem has been solved on the uplink using a new approach. The earlier known algorithms for this problem required alternating between an optimum power computation step (given the current receivers) and a SIR maximizing MSIR receiver computation step (given the current powers). In this paper the jointly optimum powers are computed independently of the MSIR receivers. Convergence has been shown analytically for not only the deterministic algorithm but, unlike in the earlier approaches, also for the blind stochastic algorithm.

References

[1] M. Varanasi, "Nonlinear Multiuser Receivers with Distributed Power Control in Cellular Radio Networks", *Proceedings of the Thirty-seventh Annual Conf. on Comm., Contr. and Computing*, pp. 820-830, Allerton, September 22-24, 1999.

[2] S. Ulukus and R. Yates, "Adaptive power Control with SIR Maximizing Multiuser Detectors," *Proc Conf. Inform. Sc. Sys.,*, March 1997.

[3] T. Guess and M. K, Varanasi, "Multiuser decision feedback receivers for the general gaussian multiple-access channel", in *Proc. 34st Annual Allerton Conf. on Comm., Contr. and Comput.*, Allerton, IL, pp. 190-199, Oct. 1996.

[4] U. Madhow and M. L. Honig, "MSIR Interference Suppression for Direct-Sequence Spread-Spectrum CDMA ", *IEEE Transcations on Communications*, vol. 38, pp. 509-519, April 1990.

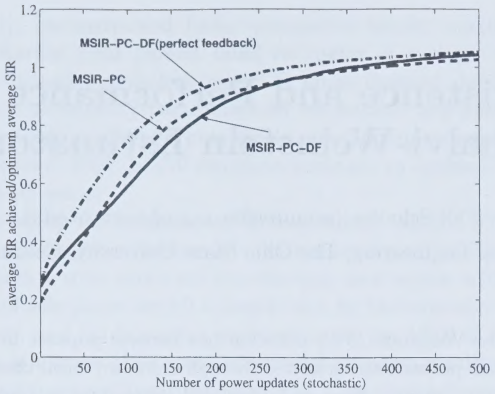


Figure 3: Average SIR achieved with the joint stochastic optimization MSIR-PC and MSIR-PC-DF (with and without perfect feedback), plotted *relative* to the common target SIR of 8.

[5] M. K. Varanasi, "Decision Feedback Multiuser Detection: A Systematic Approach, IEEE Trans. Inform. Th., vol. 45, No. 1, pp. 219-241, January 1999.

[6] E. Seneta, *Non-negative Matrices*, New York: John Wiley, 1973.

[7] B. T. Poljak, "Convergence and rate of convergence of iterative stochastic algorithms, I. General Case", *Automation and Remote Control*, No. 12, pp. 1858-1868, 1976.

[8] R. Horn and C. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, 1994.

[9] D. J. Sakrison, "Stochastic Approximation: A recursive method for solving regression problems", *Advances in Communication Systems 2*, A. V. Balakrishnan, Ed. New York: Academic, pp. 51-106, 1966.

Existence and Performance of Shalvi-Weinstein Estimators

Phil Schniter (*schniter@ee.eng.ohio-state.edu*)
Dept. of Electrical Engineering, The Ohio State University, Columbus, OH 43210.

Abstract: The Shalvi-Weinstein (SW) criterion has become popular in the design of blind linear estimators of i.i.d. processes transmitted through unknown linear channels in the presence of unknown additive interference. Here we analyze SW estimators in a general multiple-input multiple-output (MIMO) setting that allows near-arbitrary source/interference distributions and non-invertible channels. The main contributions of this paper are (i) a simple test for the existence of a SW estimator for the desired source and (ii) bounding expressions for the MSE of SW estimators that are a function of the minimum attainable MSE and the kurtoses of the source and interferers.

1 Introduction

Consider the linear estimation problem of Fig. 1, where a desired source sequence $\{s_n^{(0)}\}$ combines linearly with K interfering sources $\{s_n^{(k)}\}$ through vector channels $\{\mathbf{h}^{(0)}(z), \ldots, \mathbf{h}^{(K)}(z)\}$. Our goal is to estimate the desired source using the (vector) linear estimator $\mathbf{f}(z)$. The linear estimates $\{y_n\}$ which minimize the mean-squared error (MSE)

$$J_{\mathrm{m},\nu}(y_n) \;:=\; \mathrm{E}\{|y_n - s_{n-\nu}^{(0)}|^2\} \tag{1}$$

are generated by the minimum MSE (MMSE) estimator, or Wiener estimator, $\mathbf{f}_{\mathrm{m},\nu}(z)$. Specification of $\mathbf{f}_{\mathrm{m},\nu}(z)$, however, requires knowledge of the joint statistics of the observed sequence $\{\mathbf{r}_n\}$ and the desired source sequence $\{s_n^{(0)}\}$, which are typically unavailable when the channel is unknown.

When only the statistics of the observed sequence $\{\mathbf{r}_n\}$ are known, it may still be possible to estimate $\{s_n^{(0)}\}$ up to unknown magnitude and delay, i.e., $y_n = \sum_i \mathbf{f}_i^H \mathbf{r}_{n-i} \approx \alpha s_{n-\nu}^{(0)}$ for some $\alpha \in \mathbb{C}$, some $\nu \in \mathbb{Z}$, and all n. The literature refers to this problem as *blind* estimation (or blind deconvolution).

In [1], Shalvi and Weinstein showed that, for i.i.d. sources, noiseless invertible channels, and adequately parameterized estimators, perfect blind estimation is possible with knowledge of only the second- and fourth-order moments of the estimates $\{y_n\}$. Based on this observation, they proposed a blind estimation scheme that manipulates $\mathbf{f}(z)$ to maximize the absolute kurtosis of the estimates $\{y_n\}$ subject to a variance constraint:

$$\max\big|\mathcal{K}(y_n)\big| \quad \text{s.t.} \quad \sigma_y^2 = 1, \tag{2}$$

where, for zero-mean $\{y_n\}$, we define variance as $\sigma_y^2 := \mathrm{E}\{|y_n|^2\}$ and kurtosis as

$$\mathcal{K}(y_n) \;:=\; \mathrm{E}\{|y_n|^4\} - 2\,\mathrm{E}^2\{|y_n|^2\} - \big|\mathrm{E}\{y_n^2\}\big|^2. \tag{3}$$

As proven in [1], unconstrained linear estimators locally maximizing the Shalvi-Weinstein (SW) criterion yield perfect blind estimates of a single non-Gaussian i.i.d. source transmitted through a noiseless invertible linear channel. In practical situations, however, we expect inadequately parameterized estimators, non-invertible channels, as well as noise and/or interference of a potentially non-Gaussian nature. Are SW estimators useful in these cases? How do SW estimates compare to optimal (linear) estimators, say, in a mean-squared sense?

In this paper we study the performance of constrained ARMA SW estimators under the assumptions of the model in Section 2.1: desired source with arbitrary non-Gaussian distribution, interference with arbitrary distribution, and vector ARMA channels. The main contributions of this paper are (i) a simple test for the existence of a SW estimator for the desired source, and (ii) bounding expressions for the MSE of SW estimators that are a function of the minimum MSE attainable under the same conditions. These bounds, derived under the multi-source linear model of Fig. 1, provide a formal link between the SW and Wiener estimators in a very general context.

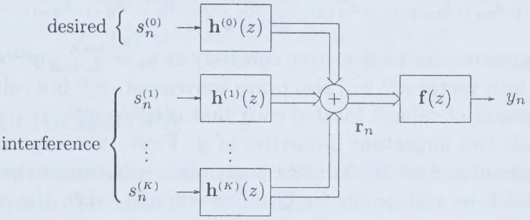


Figure 1: Linear system model with K sources of interference.

2 Background

In this section, we give more detailed information on the linear system model and the MSE, unbiased MSE, and SW criteria. The following notation is used throughout: $(\cdot)^t$ denotes transpose, $(\cdot)^*$ conjugate, $(\cdot)^H$ hermitian, and $(\cdot)^\dagger$ Moore-Penrose pseudo-inverse. Likewise, $\mathrm{E}\{\cdot\}$ denotes expectation, $\|\mathbf{x}\|_p$ the p-norm defined by $\sqrt[p]{\sum_i |x_i|^p}$, $\mathbf{I}$ the identity matrix.

2.1 Linear System Model

First we formalize the linear time-invariant multi-channel model illustrated in Fig. 1. Say that the desired symbol sequence $\{s_n^{(0)}\}$ and K sources of interference $\{s_n^{(1)}\}, \ldots, \{s_n^{(K)}\}$ each pass through separate linear "channels" before being observed at the receiver. The interference processes may correspond, e.g., to interference signals or additive noise processes. In addition, say that the receiver uses a sequence of P-dimensional vector observations $\{\mathbf{r}_n\}$ to estimate (a possibly delayed version of) the desired source sequence, where the case $P > 1$ corresponds to a receiver that employs multiple sensors and/or samples at an integer multiple of the symbol rate. The observations $\mathbf{r}_n$ can be written $\mathbf{r}_n = \sum_{k=0}^{K} \sum_{i=0}^{\infty} \mathbf{h}_i^{(k)} s_{n-i}^{(k)}$, where $\{\mathbf{h}_n^{(k)}\}$ denote the impulse response coefficients of the linear time-invariant (LTI) channel $\mathbf{h}^{(k)}(z)$. We assume that $\mathbf{h}^{(k)}(z)$ is causal and bounded-input bounded-output (BIBO) stable. Note that such $\mathbf{h}^{(k)}(z)$ admit infinite impulse response (IIR) channel models.

From the vector-valued observation sequence $\{\mathbf{r}_n\}$, the receiver generates a sequence of linear estimates $\{y_n\}$ of $\{s_{n-\nu}^{(0)}\}$, where ν is a fixed integer. Using $\{\mathbf{f}_n\}$ to denote the impulse response of the linear estimator $\mathbf{f}(z)$, the estimates are formed as $y_n = \sum_{i=-\infty}^{\infty} \mathbf{f}_i^H \mathbf{r}_{n-i}$. We will assume that the linear system $\mathbf{f}(z)$ is BIBO stable with *constrained* ARMA structure, i.e., certain polynomial coefficients in the numerator and denominator of $\mathbf{f}(z)$ may be held at zero.

In the sequel, we will focus almost exclusively on the global channel-plus-estimator response $q^{(k)}(z) := \mathbf{f}^H(z)\mathbf{h}^{(k)}(z)$. The impulse response coefficients of $q^{(k)}(z)$ can be written $q_n^{(k)} = \sum_{i=-\infty}^{\infty} \mathbf{f}_i^H \mathbf{h}_{n-i}^{(k)}$, allowing the estimates to be written as $y_n = \sum_{k=0}^{K} \sum_{i=-\infty}^{\infty} q_i^{(k)} s_{n-i}^{(k)}$. Adopting the following vector notation helps to streamline the remainder of the paper.

$$\begin{aligned}
\boldsymbol{q}^{(k)} &:= (\ldots, q_{-1}^{(k)}, q_0^{(k)}, q_1^{(k)}, \ldots)^t, \\
\boldsymbol{q} &:= (\cdots, q_{-1}^{(0)}, q_{-1}^{(1)}, \ldots, q_{-1}^{(K)}, q_0^{(0)}, q_0^{(1)}, \ldots, q_0^{(K)}, q_1^{(0)}, q_1^{(1)}, \ldots, q_1^{(K)}, \cdots)^t, \\
\boldsymbol{s}^{(k)}(n) &:= (\ldots, s_{n+1}^{(k)}, s_n^{(k)}, s_{n-1}^{(k)}, \ldots)^t, \\
\boldsymbol{s}(n) &:= (\cdots, s_{n+1}^{(0)}, s_{n+1}^{(1)}, \ldots, s_{n+1}^{(K)}, s_n^{(0)}, s_n^{(1)}, \ldots, s_n^{(K)}, s_{n-1}^{(0)}, s_{n-1}^{(1)}, \ldots, s_{n-1}^{(K)}, \cdots)^t.
\end{aligned}$$

For instance, the estimates can be rewritten concisely as $y_n = \sum_{k=0}^{K} \boldsymbol{q}^{(k)t} \boldsymbol{s}^{(k)}(n) = \boldsymbol{q}^t \boldsymbol{s}(n)$. The source-specific unit vector $\mathbf{e}_\nu^{(k)}$ will also prove convenient. $\mathbf{e}_\nu^{(k)}$ is a column vector with a single nonzero element of value 1 located such that $\boldsymbol{q}^t \mathbf{e}_\nu^{(k)} = q_\nu^{(k)}$.

We now point out two important properties of $\boldsymbol{q}$. First, it is important to recognize that a particular channel and set of estimator constraints will restrict the set of *attainable* global responses, which we will denote by $\mathcal{Q}_{\mathrm{a}}$. For example, when the estimator is FIR, $\boldsymbol{q} \in \mathcal{Q}_{\mathrm{a}} = \mathrm{row}(\boldsymbol{\mathcal{H}})$, where

$$\boldsymbol{\mathcal{H}} := \begin{pmatrix} \mathbf{h}_0^{(0)} \cdots \mathbf{h}_0^{(K)} & \mathbf{h}_1^{(0)} \cdots \mathbf{h}_1^{(K)} & \mathbf{h}_2^{(0)} \cdots \mathbf{h}_2^{(K)} & \cdots \\ 0 \;\cdots\; 0 & \mathbf{h}_0^{(0)} \cdots \mathbf{h}_0^{(K)} & \mathbf{h}_1^{(0)} \cdots \mathbf{h}_1^{(K)} & \cdots \\ \vdots \quad\;\; \vdots & \vdots \quad\;\; \vdots & \vdots \quad\;\; \vdots & \\ 0 \;\cdots\; 0 & 0 \;\cdots\; 0 & \mathbf{h}_0^{(0)} \cdots \mathbf{h}_0^{(K)} & \cdots \end{pmatrix}. \tag{4}$$

Restricting the estimator to be sparse or autoregressive, for example, would generate different attainable sets $\mathcal{Q}_{\mathrm{a}}$. Next, BIBO stable $\mathbf{f}(z)$ and $\mathbf{h}^{(k)}(z)$ imply BIBO stable $q^{(k)}(z)$, so that $\|\boldsymbol{q}^{(k)}\|_p$ exists for all $p \geq 1$, and thus $\|\boldsymbol{q}\|_p$ does as well.

Throughout the paper, we make the following assumptions on the $K+1$ source processes:

S1) For all k, $\{s_n^{(k)}\}$ is zero-mean i.i.d.

S2) $\{s_n^{(0)}\}, \ldots, \{s_n^{(K)}\}$ are jointly statistically independent.

S3) For all k, $\mathrm{E}\{|s_n^{(k)}|^2\} = \sigma_s^2$.

S4) $\mathcal{K}(s_n^{(0)}) \neq 0$.

S5) If, for any k, $q^{(k)}(z)$ or $\{s_n^{(k)}\}$ is not real-valued, then $\mathrm{E}\{s_n^{(k)2}\} = 0$ for all k.

2.2 The Unbiased Mean-Squared Error Criterion

The well-known mean-squared error (MSE) criterion was defined in (1) in terms of the estimate y_n and the estimand $s_{n-\nu}^{(0)}$. Using S1)–S3), we may rewrite (1) in terms of the global response $\boldsymbol{q}$ as follows: $J_{\mathrm{m},\nu}(\boldsymbol{q}) = \|\boldsymbol{q} - \mathbf{e}_\nu^{(0)}\|_2^2 \sigma_s^2$. Denoting MMSE quantities by the subscript "m," it can be shown that in the FIR case S1)–S3) imply that the MMSE channel-plus-estimator is given by $q_{\mathrm{m},\nu} = \mathcal{H}^t(\mathcal{H}^*\mathcal{H}^t)^\dagger\mathcal{H}^*\mathrm{e}_\nu^{(0)}$. A similar expression can be derived for the IIR case (see [2]).

Since both symbol power and channel gain are unknown in the "blind" scenario, blind estimators suffer from a gain ambiguity. To ensure that our estimator performance evaluation is meaningful in the face of such ambiguity, we base our evaluation on normalized versions of the blind estimators and normalize by the receiver gain $q_\nu^{(0)}$. Given that the estimate y_n can be decomposed into signal and interference terms as $y_n = q_\nu^{(0)} s_{n-\nu}^{(0)} + \bar{q}^t \bar{s}(n)$, where $\bar{q}$ denotes q with the $q_\nu^{(0)}$ term removed and $\bar{s}(n)$ denotes $s(n)$ with the $s_{n-\nu}^{(0)}$ term removed, the normalized estimate $y_n/q_\nu^{(0)}$ can be referred to as "conditionally unbiased" since $\mathrm{E}\{y_n/q_\nu^{(0)} | s_{n-\nu}^{(0)}\} = s_{n-\nu}^{(0)}$.

The (conditionally) unbiased MSE (UMSE) associated with y_n, an estimate of $s_{n-\nu}^{(0)}$, is then defined

$$J_{\mathrm{u},\nu}(y_n) \;:=\; \mathrm{E}\big\{|y_n/q_\nu^{(0)} - s_{n-\nu}^{(0)}|^2\big\}. \tag{5}$$

Substituting the decomposition of y_n into (5), and using S1)–S3),

$$J_{\mathrm{u},\nu}(q) \;=\; \frac{\mathrm{E}\{|\bar{q}^t\bar{s}(n)|^2\}}{|q_\nu^{(0)}|^2} \;=\; \frac{\|\bar{q}\|_2^2}{|q_\nu^{(0)}|^2}\,\sigma_s^2. \tag{6}$$

3 SW Performance

In this section we derive bounds for the UMSE of SW symbol estimators. Section 3.1 outlines our approach, Section 3.2 presents the main results, and Section 3.3 comments on these results. Proofs are not included due to lack of space, but can be found in [2].

3.1 The SW-UMSE Bounding Strategy

Since $y_n = q^t s(n)$ for $q \in \mathcal{Q}_\mathrm{a}$, source assumptions S1)-S5) imply that [3]

$$\mathcal{K}(y_n) \;=\; \sum_k \|q^{(k)}\|_4^4\, \mathcal{K}_s^{(k)}, \qquad \sigma_y^2 \;=\; \|q^{(k)}\|_2^2 \sigma_s^2, \tag{7}$$

using the shorthand $\mathcal{K}_s^{(k)} = \mathcal{K}(s_n^{(k)})$. This allows a rewrite of the SW criterion (2):

$$\max_{q \in \mathcal{Q}_\mathrm{a} \cap \mathcal{Q}_\mathrm{s}} \Big|\sum_k \|q^{(k)}\|_4^4\, \mathcal{K}_s^{(k)}\Big|$$

where $\mathcal{Q}_\mathrm{s}$ denotes the set of unit-norm global responses: $\mathcal{Q}_\mathrm{s} := \{q \text{ s.t. } \|q\|_2 = 1\}$.

Though the SW criterion admits multiple solutions, we are only interested in those that correspond to the estimation of the 0^{th} user's symbols at delay ν. We define the set of global responses *associated*[1] with the {user, delay} pair $\{0, \nu\}$ as follows:

$$\mathcal{Q}_\nu^{(0)} \;:=\; \Big\{q \text{ s.t. } |q_\nu^{(0)}| > \max_{(k,\delta)\neq(0,\nu)} |q_\delta^{(k)}|\Big\}.$$

The set of SW global responses associated with the $\{0, \nu\}$ pair is then defined by the following local maxima:

$$\{q_{\mathrm{sw},\nu}\} \;:=\; \Big\{\arg \max_{q \in \mathcal{Q}_\mathrm{a} \cap \mathcal{Q}_\mathrm{s}} \Big|\sum_k \|q^{(k)}\|_4^4\, \mathcal{K}_s^{(k)}\Big|\Big\} \cap \mathcal{Q}_\nu^{(0)}.$$

---

[1] Note that under S1)–S3), a particular user/delay combination is "associated" with an estimate if and only if that user/delay contributes more energy to the estimate than any other user/delay.

It is not possible to write general closed-form expressions for $\{\boldsymbol{q}_{\text{sw},\nu}\}$, making it difficult to characterize their performance.

Consider a reference global response $\boldsymbol{q}_{\text{r},\nu} \in \mathcal{Q}_{\text{a}} \cap \mathcal{Q}_{\text{s}} \cap \mathcal{Q}_{\nu}^{(0)}$. In other words, $\boldsymbol{q}_{\text{r},\nu}$ is an attainable unit-norm response associated with user/delay $\{0, \nu\}$. When $\boldsymbol{q}_{\text{r},\nu}$ is in the vicinity of a $\boldsymbol{q}_{\text{sw},\nu}$ (the meaning of which will be made more precise later), we know that

$$\Big|\sum_k \|\boldsymbol{q}_{\text{sw},\nu}^{(k)}\|_4^4\, \mathcal{K}_s^{(k)}\Big| \;\geq\; \Big|\sum_k \|\boldsymbol{q}_{\text{r},\nu}^{(k)}\|_4^4\, \mathcal{K}_s^{(k)}\Big| \;=\; \big|\mathcal{K}(y_{\text{r}})\big|.$$

Thus this $\boldsymbol{q}_{\text{sw},\nu}$ lies in the following set of global responses:

$$\mathcal{Q}_{\text{sw}}(\boldsymbol{q}_{\text{r},\nu}) \;:=\; \Big\{\boldsymbol{q} \;\text{ s.t. }\; \Big|\sum_k \|\boldsymbol{q}^{(k)}\|_4^4\, \mathcal{K}_s^{(k)}\Big| \geq \big|\mathcal{K}(y_{\text{r}})\big|\Big\} \cap \mathcal{Q}_{\nu}^{(0)} \cap \mathcal{Q}_{\text{s}}. \tag{8}$$

from which an SW-UMSE upper bound may be computed:

$$J_{\text{u},\nu}(\boldsymbol{q}_{\text{sw},\nu}) \;\leq\; \max_{\boldsymbol{q} \in \mathcal{Q}_{\text{sw}}(\boldsymbol{q}_{\text{r},\nu})} J_{\text{u},\nu}(\boldsymbol{q}). \tag{9}$$

Note that (9) avoids explicit consideration of the attainability constraints of $\mathcal{Q}_{\text{a}}$; they are implicitly incorporated via reference $\boldsymbol{q}_{\text{r},\nu} \in \mathcal{Q}_{\text{a}}$. Also note that the tightness of the upper bound (9) will depend on the size and shape of $\mathcal{Q}_{\text{sw}}(\boldsymbol{q}_{\text{r},\nu})$, motivating careful choice of $\boldsymbol{q}_{\text{r},\nu}$. In the sequel we choose the scaled MMSE reference $\boldsymbol{q}_{\text{r},\nu} = \boldsymbol{q}_{\text{m},\nu}/\|\boldsymbol{q}_{\text{m},\nu}\|_2$ (when $\boldsymbol{q}_{\text{m},\nu} \in \mathcal{Q}_{\nu}^{(0)}$) since it is an established benchmark with a closed-form expression.

Two simplifications will ease the evaluation of bound (9). The first is the removal of absolute value signs in the definition (8). Recognize that for $\boldsymbol{q}$ sufficiently close to $\boldsymbol{e}_{\nu}^{(0)}$, $\operatorname{sgn}\big(\sum_k \|\boldsymbol{q}^{(k)}\|_4^4\, \mathcal{K}_s^{(k)}\big) = \operatorname{sgn}\big(\mathcal{K}_s^{(k)}\big)$, in which case

$$\Big|\sum_k \|\boldsymbol{q}^{(k)}\|_4^4\, \mathcal{K}_s^{(k)}\Big| \;=\; \operatorname{sgn}\big(\mathcal{K}_s^{(k)}\big) \sum_k \|\boldsymbol{q}^{(k)}\|_4^4\, \mathcal{K}_s^{(k)}. \tag{10}$$

Our bounds will impose conditions that ensure this behavior.

Next, since both the SW and UMSE criteria are invariant to phase rotation of $\boldsymbol{q}$ (i.e., scalar multiplication of $\boldsymbol{q}$ by $e^{j\phi}$ for $\phi \in \mathbb{R}$), we can restrict the our attention to the set of "de-rotated" global responses $\{\boldsymbol{q} \text{ s.t. } q_{\nu}^{(0)} \in \mathbb{R}^+\}$. For de-rotated responses $\boldsymbol{q} \in \mathcal{Q}_{\text{s}} \cap \mathcal{Q}_{\nu}^{(0)}$, we know $q_{\nu}^{(0)} = \sqrt{1 - \|\bar{\boldsymbol{q}}\|_2^2}$, which implies that such $\boldsymbol{q}$ are completely described by their interference response $\bar{\boldsymbol{q}}$ (as described in Section 2.2). Moreover, these interference responses lie within $\bar{\mathcal{Q}}_{\nu}^{(0)}$, the projection of $\mathcal{Q}_{\nu}^{(0)} \cap \mathcal{Q}_{\text{s}}$ onto $\{\bar{\boldsymbol{q}}\}$:

$$\bar{\mathcal{Q}}_{\nu}^{(0)} \;:=\; \Big\{\bar{\boldsymbol{q}} \text{ s.t. } \sqrt{1 - \|\bar{\boldsymbol{q}}\|_2^2} > \max_{(k,\delta) \neq (0,\nu)} |q_{\delta}^{(k)}|\Big\}.$$

(See Fig. 2 for the construction of $\bar{\mathcal{Q}}_{\nu}^{(0)}$, whose boundary is illustrated by the thick shaded curves.) Using this parameterization, (6) and (7) imply

$$\begin{aligned}
J_{\text{u},\nu}(\boldsymbol{q}_{\text{sw},\nu})/\sigma_s^2\Big|_{\boldsymbol{q} \in \mathcal{Q}_{\text{s}} \cap \mathcal{Q}_{\nu}^{(0)}} &= \|\bar{\boldsymbol{q}}\|_2^2 / (1 - \|\bar{\boldsymbol{q}}\|_2^2) \\
\sum_k \|\boldsymbol{q}^{(k)}\|_4^4\, \mathcal{K}_s^{(k)}\Big|_{\boldsymbol{q} \in \mathcal{Q}_{\text{s}} \cap \mathcal{Q}_{\nu}^{(0)}} &= \big(1 - \|\bar{\boldsymbol{q}}\|_2^2\big)^2 \mathcal{K}_s^{(0)} + \sum_k \|\bar{\boldsymbol{q}}^{(k)}\|_4^4\, \mathcal{K}_s^{(k)}.
\end{aligned} \tag{11}$$

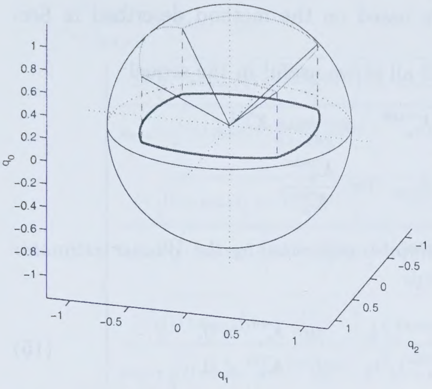
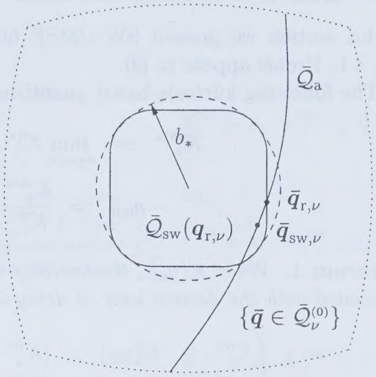


Figure 2: $\bar{\mathcal{Q}}_\nu^{(0)}$, created by projecting $\mathcal{Q}_\nu^{(0)} \cap \mathcal{Q}_\mathrm{s}$ onto the interference space. The boundary of $\bar{\mathcal{Q}}_\nu^{(0)}$ is demarcated by the thick shaded curves.



$\bar{\mathcal{Q}}_\mathrm{a}$
b_*
$\bar{\boldsymbol{q}}_{\mathrm{r},\nu}$
$\bar{\mathcal{Q}}_\mathrm{sw}(\boldsymbol{q}_{\mathrm{r},\nu})$
$\bar{\boldsymbol{q}}_{\mathrm{sw},\nu}$
$\{\bar{\boldsymbol{q}} \in \bar{\mathcal{Q}}_\nu^{(0)}\}$

Figure 3: Illustration of SW-UMSE bounding technique in the interference space $\{\bar{\boldsymbol{q}}\}$, i.e., the horizontal plane in Fig. 2.

With the two simplifications above, (9) becomes

$$J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{sw},\nu}) \;\le\; \max_{\bar{\boldsymbol{q}} \in \bar{\mathcal{Q}}_\mathrm{sw}(\boldsymbol{q}_{\mathrm{r},\nu})} J_{\mathrm{u},\nu}(\bar{\boldsymbol{q}}).$$

where $\bar{\mathcal{Q}}_\mathrm{sw}$ is the following $\{\bar{\boldsymbol{q}}\}$-space projection of $\mathcal{Q}_\mathrm{sw}$:

$$\bar{\mathcal{Q}}_\mathrm{sw}(\boldsymbol{q}_{\mathrm{r},\nu}) := \begin{cases} \left\{\bar{\boldsymbol{q}} \in \bar{\mathcal{Q}}_\nu^{(0)} \text{ s.t. } \left(1-\|\bar{\boldsymbol{q}}\|_2^2\right)^2 \mathcal{K}_s^{(0)} + \sum_k \|\bar{\boldsymbol{q}}^{(k)}\|_4^4 \,\mathcal{K}_s^{(k)} \ge \mathcal{K}(y_\mathrm{r})\right\}, & \mathcal{K}_s^{(0)} > 0, \\ \left\{\bar{\boldsymbol{q}} \in \bar{\mathcal{Q}}_\nu^{(0)} \text{ s.t. } \left(1-\|\bar{\boldsymbol{q}}\|_2^2\right)^2 \mathcal{K}_s^{(0)} + \sum_k \|\bar{\boldsymbol{q}}^{(k)}\|_4^4 \,\mathcal{K}_s^{(k)} \le \mathcal{K}(y_\mathrm{r})\right\}, & \mathcal{K}_s^{(0)} < 0. \end{cases} \tag{12}$$

Finally, since $J_{\mathrm{u},\nu}(\bar{\boldsymbol{q}})$ is strictly increasing in $\|\bar{\boldsymbol{q}}\|_2$ (over its valid range), we claim

$$J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{sw},\nu}) \;\le\; b_*^2/(1-b_*^2) \quad \text{where} \quad b_* \;:=\; \max_{\bar{\boldsymbol{q}} \in \bar{\mathcal{Q}}_\mathrm{sw}(\boldsymbol{q}_{\mathrm{r},\nu})} \|\bar{\boldsymbol{q}}\|_2. \tag{13}$$

The constrained maximization of b_* can be restated as the following minimization.

$$b_* \;=\; \min_b \text{ s.t. } \left\{\bar{\boldsymbol{q}} \in \bar{\mathcal{Q}}_\mathrm{sw}(\boldsymbol{q}_{\mathrm{r},\nu}) \Rightarrow \|\bar{\boldsymbol{q}}\|_2 \le b\right\} \tag{14}$$

Fig. 3 presents a summary of the bounding procedure in the interference response space $\{\bar{\boldsymbol{q}}\}$. The set of attainable interference responses is denoted by $\bar{\mathcal{Q}}_\mathrm{a}$, which can be interpreted as a projection of $\mathcal{Q}_\mathrm{a} \cap \mathcal{Q}_\mathrm{s} \cap \mathcal{Q}_\nu^{(0)}$ onto $\{\bar{\boldsymbol{q}}\}$. Notice that the reference response $\bar{\boldsymbol{q}}_{\mathrm{r},\nu}$ and the SW response $\bar{\boldsymbol{q}}_{\mathrm{sw},\nu}$ both lie on $\bar{\mathcal{Q}}_\mathrm{a}$. Though the exact location of $\bar{\boldsymbol{q}}_{\mathrm{sw},\nu}$ is unknown, we know that it is contained by $\bar{\mathcal{Q}}_\mathrm{sw}(\boldsymbol{q}_{\mathrm{r},\nu})$, depicted in Fig. 3 by the shaded region. Thus, an upper bound on the UMSE of the SW estimator can be calculated using b_*, the maximum interference radius over $\bar{\mathcal{Q}}_\mathrm{sw}(\boldsymbol{q}_{\mathrm{r},\nu})$. As a cautionary note, there exist situations where the shape of $\bar{\mathcal{Q}}_\mathrm{sw}(\boldsymbol{q}_{\mathrm{r},\nu})$ prevents containment by a $\bar{\boldsymbol{q}}$-space ball. In the next section we present conditions (derived in [2]) which avoid these problematic situations.

3.2 The SW-UMSE Bounds

In this section we present SW-UMSE bounds based on the method described in Section 3.1. Proofs appear in [2].

The following kurtosis-based quantities will all prove useful in the sequel:

$$\begin{aligned} \mathcal{K}_s^{\min} &:= \min_{0\le k\le K} \mathcal{K}_s^{(k)}, & \mathcal{K}_s^{\max} &:= \max_{0\le k\le K} \mathcal{K}_s^{(k)} \\ \rho_{\min} &:= \frac{\mathcal{K}_s^{\min}}{\mathcal{K}_s^{(0)}}, & \rho_{\max} &:= \frac{\mathcal{K}_s^{\max}}{\mathcal{K}_s^{(0)}} \end{aligned}$$

Theorem 1. *When $\mathcal{K}(y_{\mathrm{m}})$, the kurtosis of estimates generated by the Wiener estimator associated with the desired user at delay ν, obeys*

$$\begin{cases} \mathcal{K}_s^{(0)} \ \ge\ \mathcal{K}(y_{\mathrm{m}}) \ >\ \big(\mathcal{K}_s^{(0)} + \mathcal{K}_s^{\max}\big)/4, & \textit{for } \mathcal{K}_s^{(0)} > 0, \\ \mathcal{K}_s^{(0)} \ \le\ \mathcal{K}(y_{\mathrm{m}}) \ <\ \big(\mathcal{K}_s^{(0)} + \mathcal{K}_s^{\min}\big)/4, & \textit{for } \mathcal{K}_s^{(0)} < 0, \end{cases} \tag{15}$$

the UMSE of SW estimators associated with the same user/delay can be upper bounded by $J_{\mathrm{u},\nu}\big|_{\mathrm{sw},\nu}^{\max,\mathcal{K}(y_{\mathrm{m}})}$, where

$$J_{\mathrm{u},\nu}\Big|_{\mathrm{sw},\nu}^{\max,\mathcal{K}(y_{\mathrm{m}})} := \begin{cases} \dfrac{1-\sqrt{(\rho_{\max}+1)\frac{\mathcal{K}(y_{\mathrm{m}})}{\mathcal{K}_s^{(0)}}-\rho_{\max}}}{\rho_{\max}+\sqrt{(\rho_{\max}+1)\frac{\mathcal{K}(y_{\mathrm{m}})}{\mathcal{K}_s^{(0)}}-\rho_{\max}}}\,\sigma_s^2, & \textit{for } \mathcal{K}_s^{(0)} > 0, \\[2ex] \dfrac{1-\sqrt{(\rho_{\min}+1)\frac{\mathcal{K}(y_{\mathrm{m}})}{\mathcal{K}_s^{(0)}}-\rho_{\min}}}{\rho_{\min}+\sqrt{(\rho_{\min}+1)\frac{\mathcal{K}(y_{\mathrm{m}})}{\mathcal{K}_s^{(0)}}-\rho_{\min}}}\,\sigma_s^2, & \textit{for } \mathcal{K}_s^{(0)} < 0. \end{cases} \tag{16}$$

Furthermore, (15) guarantees the existence of a SW estimator associated with this user/delay when $\boldsymbol{q}$ is FIR.

While Theorem 1 presents a closed-form SW-UMSE bounding expression in terms of the kurtosis of the MMSE estimates, it is also possible to derive lower and upper bounds in terms of the UMSE of the MMSE estimator.

Theorem 2. *If $J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu}) < J_{\mathrm{o}}\sigma_s^2$, where*

$$J_{\mathrm{o}} := \begin{cases} 2\sqrt{(1+\rho_{\max})^{-1}}-1 & \mathcal{K}_s^{(0)} > 0,\ \mathcal{K}_s^{\min} \ge 0 \\ \frac{1-\sqrt{1-(3-\rho_{\max})(1+\rho_{\min})/4}}{\rho_{\min}+\sqrt{1-(3-\rho_{\max})(1+\rho_{\min})/4}}, & \mathcal{K}_s^{(0)} > 0,\ \mathcal{K}_s^{\min} < 0,\ \mathcal{K}_s^{\min} \ne -\mathcal{K}_s^{(0)} \\ \frac{3-\rho_{\max}}{5+\rho_{\max}} & \mathcal{K}_s^{(0)} > 0,\ \mathcal{K}_s^{\min} < 0,\ \mathcal{K}_s^{\min} = -\mathcal{K}_s^{(0)} \\ 2\sqrt{(1+\rho_{\min})^{-1}}-1 & \mathcal{K}_s^{(0)} < 0,\ \mathcal{K}_s^{\max} \le 0 \\ \frac{1-\sqrt{1-(3-\rho_{\min})(1+\rho_{\max})/4}}{\rho_{\max}+\sqrt{1-(3-\rho_{\min})(1+\rho_{\max})/4}}, & \mathcal{K}_s^{(0)} < 0,\ \mathcal{K}_s^{\max} > 0,\ \mathcal{K}_s^{\max} \ne -\mathcal{K}_s^{(0)} \\ \frac{3-\rho_{\min}}{5+\rho_{\min}} & \mathcal{K}_s^{(0)} < 0,\ \mathcal{K}_s^{\max} > 0,\ \mathcal{K}_s^{\max} = -\mathcal{K}_s^{(0)} \end{cases} \tag{17}$$

the UMSE of SW estimators associated with the same user/delay can be bounded as follows:

$$J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu}) \ \le\ J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{sw},\nu}) \ \le\ J_{\mathrm{u},\nu}\Big|_{\mathrm{sw},\nu}^{\max,\mathcal{K}(y_{\mathrm{m}})} \ \le\ J_{\mathrm{u},\nu}\Big|_{\mathrm{sw},\nu}^{\max,J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu})},$$

where

$$J_{\mathrm{u},\nu}\Big|_{\mathrm{sw},\nu}^{\max,J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu})} := \tag{18}$$

$$\begin{cases} \dfrac{1-\sqrt{(1+\rho_{\max})\left(1+\frac{J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu})}{\sigma_s^2}\right)^{-2}-\rho_{\max}}}{\rho_{\max}+\sqrt{(1+\rho_{\max})\left(1+\frac{J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu})}{\sigma_s^2}\right)^{-2}-\rho_{\max}}}\,\sigma_s^2 & \mathcal{K}_s^{(0)}>0,\ \mathcal{K}_s^{\min}\ge 0 \\[3ex] \dfrac{1-\sqrt{(1+\rho_{\max})\left(1+\frac{J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu})}{\sigma_s^2}\right)^{-2}\left(1+\rho_{\min}\frac{J_{\mathrm{u},\nu}^2(\boldsymbol{q}_{\mathrm{m},\nu})}{\sigma_s^4}\right)-\rho_{\max}}}{\rho_{\max}+\sqrt{(1+\rho_{\max})\left(1+\frac{J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu})}{\sigma_s^2}\right)^{-2}\left(1+\rho_{\min}\frac{J_{\mathrm{u},\nu}^2(\boldsymbol{q}_{\mathrm{m},\nu})}{\sigma_s^4}\right)-\rho_{\max}}}\,\sigma_s^2 & \mathcal{K}_s^{(0)}>0,\ \mathcal{K}_s^{\min}<0 \\[3ex] \dfrac{1-\sqrt{(1+\rho_{\min})\left(1+\frac{J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu})}{\sigma_s^2}\right)^{-2}-\rho_{\min}}}{\rho_{\min}+\sqrt{(1+\rho_{\min})\left(1+\frac{J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu})}{\sigma_s^2}\right)^{-2}-\rho_{\min}}}\,\sigma_s^2 & \mathcal{K}_s^{(0)}<0,\ \mathcal{K}_s^{\max}\le 0 \\[3ex] \dfrac{1-\sqrt{(1+\rho_{\min})\left(1+\frac{J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu})}{\sigma_s^2}\right)^{-2}\left(1+\rho_{\max}\frac{J_{\mathrm{u},\nu}^2(\boldsymbol{q}_{\mathrm{m},\nu})}{\sigma_s^4}\right)-\rho_{\min}}}{\rho_{\min}+\sqrt{(1+\rho_{\min})\left(1+\frac{J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu})}{\sigma_s^2}\right)^{-2}\left(1+\rho_{\max}\frac{J_{\mathrm{u},\nu}^2(\boldsymbol{q}_{\mathrm{m},\nu})}{\sigma_s^4}\right)-\rho_{\min}}}\,\sigma_s^2 & \mathcal{K}_s^{(0)}<0,\ \mathcal{K}_s^{\max}>0 \end{cases}$$

Furthermore, (17) guarantees the existence of a SW estimator associated with this user/delay when $\boldsymbol{q}$ is FIR.

Equation (18) leads to an elegant approximation of the *extra* UMSE of SW estimators:

$$\mathcal{E}_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{sw},\nu}) := J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{sw},\nu}) - J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu}).$$

Theorem 3. *If $J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu}) < J_{\mathrm{o}}\sigma_s^2$, then the extra UMSE of SW estimators can be bounded as $\mathcal{E}_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{sw},\nu}) \le \mathcal{E}_{\mathrm{u},\nu}\Big|_{\mathrm{c},\nu}^{\max,J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu})}$, where*

$$\begin{aligned} \mathcal{E}_{\mathrm{u},\nu}\Big|_{\mathrm{c},\nu}^{\max,J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu})} &:= J_{\mathrm{u},\nu}\Big|_{\mathrm{sw},\nu}^{\max,J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu})} - J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu}) \\ &= \begin{cases} \frac{1}{2\sigma_s^2}\rho_{\max} J_{\mathrm{u},\nu}^2(\boldsymbol{q}_{\mathrm{m},\nu}) + \mathcal{O}\big(J_{\mathrm{u},\nu}^3(\boldsymbol{q}_{\mathrm{m},\nu})\big) & \mathcal{K}_s^{(0)}>0,\ \mathcal{K}_s^{\max}\ge 0 \\ \frac{1}{2\sigma_s^2}(\rho_{\max}-\rho_{\min}) J_{\mathrm{u},\nu}^2(\boldsymbol{q}_{\mathrm{m},\nu}) + \mathcal{O}\big(J_{\mathrm{u},\nu}^3(\boldsymbol{q}_{\mathrm{m},\nu})\big) & \mathcal{K}_s^{(0)}>0,\ \mathcal{K}_s^{\max}<0 \\ \frac{1}{2\sigma_s^2}\rho_{\min} J_{\mathrm{u},\nu}^2(\boldsymbol{q}_{\mathrm{m},\nu}) + \mathcal{O}\big(J_{\mathrm{u},\nu}^3(\boldsymbol{q}_{\mathrm{m},\nu})\big) & \mathcal{K}_s^{(0)}<0,\ \mathcal{K}_s^{\max}\le 0 \\ \frac{1}{2\sigma_s^2}(\rho_{\min}-\rho_{\max}) J_{\mathrm{u},\nu}^2(\boldsymbol{q}_{\mathrm{m},\nu}) + \mathcal{O}\big(J_{\mathrm{u},\nu}^3(\boldsymbol{q}_{\mathrm{m},\nu})\big) & \mathcal{K}_s^{(0)}<0,\ \mathcal{K}_s^{\max}>0 \end{cases} \end{aligned} \tag{19}$$

Equation (19) implies that the extra UMSE of SW estimators is upper bounded by approximately the *square* of the minimum UMSE. Fig. 4 plots the upper bound on SW-UMSE and extra SW-UMSE from (18) as a function of $J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu})/\sigma_s^2$ for various values of $\rho_{\min}$ and $\rho_{\max}$. The second-order approximation based on (19) appears very good for all but the largest values of UMSE.

3.3 Comment: Implicit Incorporation of $\mathcal{Q}_{\mathrm{a}}$

First, recall that the SW-UMSE bounding procedure incorporated $\mathcal{Q}_{\mathrm{a}}$, the set of attainable global responses, *only* in the requirement that $\boldsymbol{q}_{\mathrm{r},\nu} \in \mathcal{Q}_{\mathrm{a}} \cap \mathcal{Q}_{\mathrm{s}} \cap \mathcal{Q}_\nu^{(0)}$. Thus

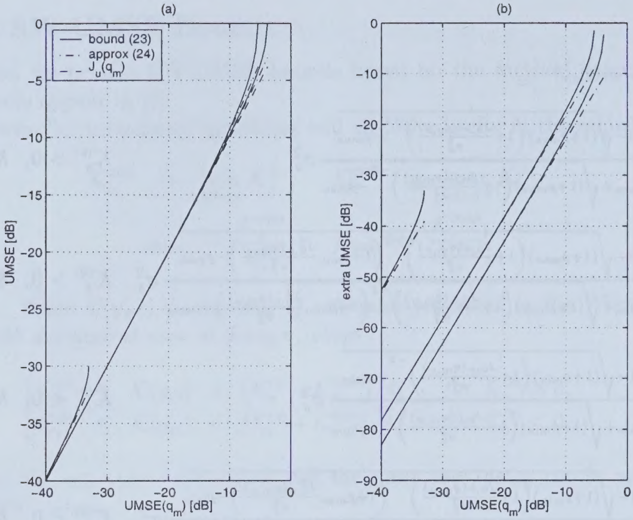


Figure 4: Upper bound on (a) SW-UMSE and (b) extra SW-UMSE versus $J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu})$ (when $\sigma_s^2 = 1$) from (18) with second-order approximation from (19). From left to right, $\{\rho_{\min}, \rho_{\max}\} = \{1000, 0\}$, $\{1, -2\}$, and $\{1, 0\}$.

Theorems 1–3, written under the reference choice $\boldsymbol{q}_{\mathrm{r},\nu} = \boldsymbol{q}_{\mathrm{m},\nu}/\|\boldsymbol{q}_{\mathrm{m},\nu}\|_2 \in \mathcal{Q}_{\mathrm{a}} \cap \mathcal{Q}_{\mathrm{s}} \cap \mathcal{Q}_\nu^{(0)}$, implicitly incorporate the channel and/or estimator constraints that define $\mathcal{Q}_{\mathrm{a}}$. For example, if $\boldsymbol{q}_{\mathrm{m},\nu}$ is the MMSE response constrained to the set of causal IIR estimators, then SW-UMSE bounds based on this $\boldsymbol{q}_{\mathrm{m},\nu}$ will implicitly incorporate the causality constraint. The implicit incorporation of the attainable set $\mathcal{Q}_{\mathrm{a}}$ makes these bounding theorems quite general and easy to use.

4 Numerical Examples

Here we present the results of an experiment which compares our UMSE bounds to the UMSE characterizing SW estimators found by gradient descent.[2] Ten super-Gaussian sources ($\mathcal{K}_s^{(k)} = 2$) were mixed using matrix $\mathcal{H}$ with real-valued zero-mean Gaussian entries. The estimator observed the mixture in the presence of AWGN (at -20dB) and generated estimates of a particular source using 8 adjustable parameters. Note that the number of sensors is less than the number of sources and that noise is present, implying that $\mathcal{H}$ is not full column rank and perfect estimation is not possible.

Fig. 5(a) plots the UMSE upper bounds $J_{\mathrm{u},\nu}\big|_{\mathrm{sw},\nu}^{\max,\mathcal{K}(y_{\mathrm{m}})}$ and $J_{\mathrm{u},\nu}\big|_{\mathrm{sw},\nu}^{\max,J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu})}$ for comparison with $J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{sw},\nu})$. As a means of "zooming in" on the small differences in UMSE, Fig. 5(b) plots the extra-UMSE upper bounds $\mathcal{E}_{\mathrm{u},\nu}\big|_{\mathrm{c},\nu}^{\max,\mathcal{K}(\boldsymbol{q}_{\mathrm{m},\nu})}$ and $\mathcal{E}_{\mathrm{u},\nu}\big|_{\mathrm{c},\nu}^{\max,J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu})}$. The $J_{\mathrm{u},\nu}(\boldsymbol{q}_{\mathrm{m},\nu})$-based bounds are denoted by solid lines, the $\mathcal{K}(\boldsymbol{q}_{\mathrm{m},\nu})$-based bounds are denoted by $\bullet$'s, and the gradient-descent values are denoted by $\times$'s. Note the tightness of the bounds for all but the largest values of $J_{u,\nu}(\boldsymbol{q}_{m,\nu})$.

---

[2] Gradient descent results were obtained by the MATLAB routine "`fmincon`," which was initialized randomly in a small ball around the MMSE estimator.

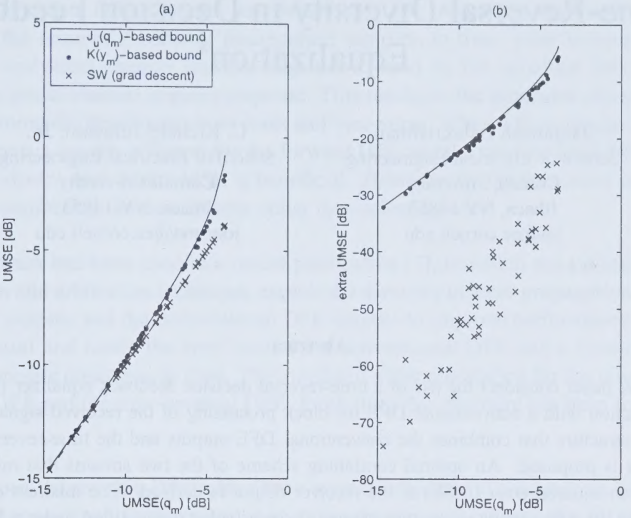


Figure 5: Bounds on SW-UMSE for $N_f = 8$, 10 sources with $\mathcal{K}_s^{(k)} = 2$, AWGN at -20dB, and random $\mathcal{H}$.

5 Conclusions

In this paper we have derived conditions under which SW estimators exist and derived bounds for the UMSE of SW estimators. The existence conditions are simple tests which guarantee a SW estimator for the desired user at a particular delay, and these existence arguments have been proven for vector-valued FIR channels and constrained vector-valued FIR estimators. The first bound is a function of the kurtosis of the MMSE estimates, while the second bound is a function of the minimum UMSE. Analysis of the second bound shows that the extra UMSE of SW estimators is upper bounded by approximately the square of the minimum UMSE. Thus, SW estimators are very close (in a MSE sense) to optimum linear estimators when the minimum MSE is small. Numerical simulations suggest that the bounds are reasonably tight.

References

[1] O. Shalvi and E. Weinstein, "New criteria for blind deconvolution of nonminimum phase systems (channels)," *IEEE Trans. on Info. Theory*, vol. 36, no. 2, pp. 312-21, Mar. 1990.

[2] P. Schniter, "Blind Estimation Without Priors: Performance, Convergence, and Efficient Implementation," *Ph.D. dissertation*, Cornell University, Ithaca, NY, May 2000.

[3] B. Porat, *Digital Processing of Random Signals*, Englewood Cliffs, NJ: Prentice-Hall, 1994.

Time-Reversal Diversity in Decision Feedback Equalization*

Jaiganesh Balakrishnan
School of Electrical Engineering
Cornell University
Ithaca, NY 14853
jai@ee.cornell.edu

C. Richard Johnson, Jr.,
School of Electrical Engineering
Cornell University
Ithaca, NY 14853
johnson@ee.cornell.edu

Abstract

This paper considers the use of a time-reversal decision feedback equalizer (DFE) in conjunction with a conventional DFE for block processing of the received signal. A receiver structure that combines the conventional DFE outputs and the time-reversal DFE outputs is proposed. An optimal combining scheme of the two streams that minimizes the mean-squared error (MSE) at the receiver output is derived. The inherent diversity between the soft outputs of the two streams is described and quantified under a few simplifying assumptions. For the special case of a symmetric channel, the optimal DFE filter setting that minimizes the MSE at the output of the diversity combining block is derived. Simulation results demonstrating performance gains for the proposed receiver structure are presented for both symmetric and asymmetric channels.

1 Introduction

Future generation wireless communication technology promises an increase in data rate capabilities of present radio links. As we go for higher rates the frequency distortion caused by the channel becomes worse. The distortion caused by the resulting overlap of received symbols is called intersymbol interference (ISI), which, if left uncompensated, causes high error rates. One suboptimal scheme that is widely used in practice, due to its simple implementation, is the linear equalizer. But, in suppressing the intersymbol interference, the linear equalizer inevitably enhances the noise. This basic limitation of a linear equalizer's ability to cope with severe ISI has motivated a considerable amount of research into suboptimal nonlinear equalizers with low computational complexity such as the decision feedback equalizer (DFE).

The idea of using previous decisions to cope with the ISI problem was first introduced in 1967 by Austin [1]. The decision feedback receiver that minimizes the mean-squared error (MSE) between the input to the threshold device and the transmitted symbol was first obtained by Monsen [2]. The action of the DFE is to feed back a weighted sum of the past decisions, assumed to be correct, to cancel the ISI they cause in the present signal interval. However, when incorrect decisions are fed back through the feedback filter of a DFE, it induces a burst

---

*This work was supported by NSF Grants EEC - 9872436 and ECS - 9811297, Applied Signal Technology and Lucent Technologies.

of errors causing performance degradation and this phenomenon is known as error propagation.

Bidirectional equalization of the received signal was first considered in [3, 4]. The use of a selective time-reversal DFE structure was described in [5]. A time-reversal operation is done by reversing the sequential order of the received samples, in time, prior to equalization. As a result, the equivalent channel impulse response as seen by the equalizer becomes a time-reverse of the actual channel impulse response. This results in the minimum phase roots of the channel becoming the maximum phase roots and vice-versa. When a finite length DFE is used, the performance is usually different for the forward DFE and the time-reversal DFE and hence selecting the stream with lesser MSE is beneficial. However, the performance can be further improved by combining the two streams rather than selecting one.

This approach has been used in a recent publication [7], in which the authors, based on a reconstruction and arbitration technique, exploit the diversity in error propagation between the forward DFE outputs and the time-reversal DFE outputs to improve performance. Error propagation is causal and hence the error burst for a conventional DFE and a time-reversal DFE proceed in opposite directions in time. This results in a low correlation for the errors caused by the forward DFE and the time-reversal DFE. Exploiting this diversity has been the motivation behind [7].

In this paper, the use of a DFE with a time-reversal structure in conjunction with a conventional DFE is considered. Furthermore, it is shown that the soft outputs of a forward DFE and the time-reversal DFE possess an inherent diversity. The term diversity is used here to refer to the low correlation in the error terms of the soft outputs. The diversity is a result of the nonlinear structure of the DFE and the fact that we allow anti-causal processing of the received signal and is present even in the absence of error propagation. This diversity, when exploited, results in a decrease in the noise gain term and the residual ISI component. A receiver structure that combines the conventional DFE output and the time-reversal DFE output, to exploit this inherent diversity between the two streams, is proposed and the diversity that can be potentially exploited is quantified.

The organization of the paper is as follows. Section 2 describes the system model and states the assumptions made in this paper. An optimal combining scheme for the two streams that minimizes the MSE at the output of the receiver structure is derived in Section 3. Section 4 describes the time-reversal diversity inherent in the noise gain term and the residual ISI component for a symmetric channel. In Section 5, the optimal unbiased DFE tap setting that minimizes the MSE at the output of the diversity combining scheme is derived for a symmetric channel. Simulation results demonstrating performance gains of the proposed receiver structure are provided in Section 6 for both symmetric and asymmetric channels. A summary of the contributions of this paper and a brief discussion of possible directions for future work is provided in Section 7.

2 System Model

The block diagram of the proposed receiver structure is shown in Figure 1. The source symbols $s(n)$ are transmitted though a frequency selective channel. The channel incorporates the effects of frequency selective distortion caused by the physical medium and transmit pulse shaping and is modeled as a linear FIR filter $\mathbf{c}$. The transmitted signal is further corrupted by additive white

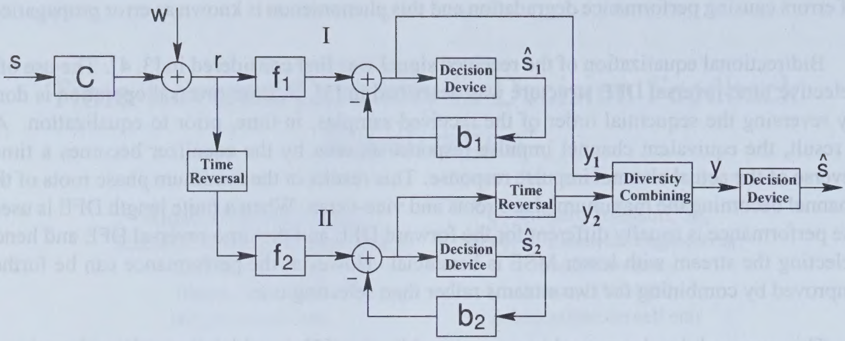


Figure 1: Time-Reversal DFE structure

noise at the receiver. The received signal $r(n)$ is given by,

$$r(n)=\sum_{k=0}^{L_c-1} c(k)s(n-k)+w(n), \quad n=1,2,\cdots,N \tag{1}$$

where $\mathbf{c}=[c(0)\ c(1)\ \ldots\ c(L_c-1)]^T$ is the channel impulse response with L_c taps and w(n) is the noise term.

The receiver structure of Figure 1 processes a data block of size N using two different streams. The received signal $r(n)$ is equalized in stream I using a DFE with an inherent detection delay $\delta 1$. The feedforward filter $\mathbf{f_1}$ has L_{f1} taps and its impulse response is denoted by the vector $\mathbf{f_1}=[f_1(0)\ f_1(1)\ \ldots\ f_1(L_{f1}-1)]^T$ while the feedback filter $\mathbf{b_1}$ has L_{b1} taps and an impulse response $\mathbf{b_1}=[b_1(1)\ b_1(2)\ \ldots\ b_1(L_{b1})]^T$. The combined channel-feedforward filter impulse response for stream I is given by $\mathbf{h_1}=\mathbf{c}*\mathbf{f_1}$, where $*$ denotes the convolution operation. $h_1(\delta 1)$ is then referred to as the cursor.

The received signal $r(n)$ is time reversed in stream II. The time-reversed version of the received signal can then be expressed as

$$\begin{aligned}\tilde{r}(n) &= r(N-n+1)\\ &= \sum_{k=-(L_c-1)}^{0} c(-k)\tilde{s}(n-k)+\tilde{w}(n)\end{aligned} \tag{2}$$

where tilde implies time-reversal, $\tilde{s}(n)=s(N-n+1)$ is the time-reversed source sequence and $\tilde{w}(n)=w(N-n+1)$ is the time-reversed noise sequence. This is equivalent to transmitting the time-reversed source sequence through a channel whose coefficients are the time-reversed version of c. Let $\mathbf{f_2}$ with L_{f2} taps and $\mathbf{b_2}$ with L_{b2} taps be the feedforward and feedback filter settings of the DFE that equalizes the time-reversed channel $\tilde{\mathbf{c}}=[c(L_c-1)\ \cdots\ c(1)\ c(0)]^T$. The combined channel-feedforward filter impulse response for this stream is $\mathbf{h_2}=\tilde{\mathbf{c}}*\mathbf{f_2}$. Let $\delta 2$ be the delay introduced by the equalizer in stream II. Note that the optimum burst synchronization for the two streams may be different (see [5]), in general, as it depends on the detection delay of the equalizer. For a particular equalizer structure we may explicitly include a time shift by a multiple of the symbol duration in equation (2) to account for such a change in receiver timing for the two modes of operation.

In this paper we make the following assumptions.

- The FFF and the FBF for the two streams are those designed to generate the unbiased DFE settings, i.e., $h_1(\delta 1) = h_2(\delta 2) = 1$.
- The decisions that are fed back to the FBF of the DFE are correct, i.e., ideal feedback.
- The source sequence $s(n)$ is i.i.d and has unit variance.
- The noise sequence $w(n)$ is zero mean, white, has a variance σ_w^2 and is uncorrelated to the source sequence.

$y_1(n)$ are the soft inputs to the decision device of stream I and $y_2(n)$ are the time-reversed soft inputs to the decision device of stream II. The soft input sequences $y_1(n)$ and $y_2(n)$ can be expressed as,

$$y_1(n) = s(n) + \eta_1(n) \tag{3}$$

$$y_2(n) = s(n) + \eta_2(n) \tag{4}$$

where

$$\begin{aligned}\eta_1(n) = &\sum_{m=0}^{L_{f1}-1} f_1(m)w(n+\delta 1-m) + \sum_{m=0}^{\delta 1-1} h_1(m)s(n+\delta 1-m)\\ &+ \sum_{m=\delta 1+L_{b1}+1}^{L_{f1}+L_c-2} h_1(m)s(n+\delta 1-m)\end{aligned} \tag{5}$$

$$\begin{aligned}\eta_2(n) = &\sum_{m=0}^{L_{f2}-1} f_2(m)w(n+L_c-1-\delta 2+m) + \sum_{m=0}^{\delta 2-1} h_2(m)s(n-\delta 2+m)\\ &+ \sum_{m=\delta 2+L_{b2}+1}^{L_{f2}+L_c-2} h_2(m)s(n-\delta 2+m)\end{aligned} \tag{6}$$

Let σ_1^2 and σ_2^2 be the variances of $\eta_1(\cdot)$ and $\eta_2(\cdot)$ respectively. We use ρ to denote the coefficient of correlation between the sequences $\eta_1(\cdot)$ and $\eta_2(\cdot)$, i.e.,

$$\rho = \frac{\mathbf{E}[\eta_1\eta_2]}{\sigma_1\sigma_2} \tag{7}$$

where $\mathbf{E}[\cdot]$ is the expectation operation.

3 Diversity Combining

The soft output $y(n)$ of the diversity combining block in Figure 1 is generated as a linear combination of the sequences $y_1(n)$ and $y_2(n)$, i.e.,

$$\begin{aligned} y(n) &= \alpha y_1(n) + (1-\alpha)y_2(n) \\ &= s(n) + \alpha\eta_1(n) + (1-\alpha)\eta_2(n) \end{aligned} \tag{8}$$

The MSE for the soft output $y(n)$ is given by,

$$\begin{aligned} \text{MSE} &= \mathbf{E}[|y(n) - s(n)|^2] \\ &= \alpha^2\sigma_1^2 + (1-\alpha)^2\sigma_2^2 + 2\rho\alpha(1-\alpha)\sigma_1\sigma_2 \end{aligned} \tag{9}$$

The weighting factor α is optimized to minimize MSE and the optimal value of α is obtained by setting,

$$\frac{\partial \text{MSE}}{\partial \alpha} = 0 \tag{10}$$

$$\Rightarrow \qquad \alpha = \frac{\sigma_2^2 - \rho\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2} \tag{11}$$

When $\sigma_1 = \sigma_2$, the optimal weighting factor is $\alpha = \frac{1}{2}$ and this is known as equal-gain combining.

4 Time-Reversal Diversity

For a symmetric channel, the time-reversed channel impulse $\tilde{\mathbf{c}}$ is equal to the channel impulse response $\mathbf{c}$. Hence, the same DFE tap settings can be used for the two streams, i.e., $\mathbf{f_1} = \mathbf{f_2} = \mathbf{f}$ and $\mathbf{b_1} = \mathbf{b_2} = \mathbf{f}$. Further, the mean-squared error (MSE) for the two streams, σ_1 and σ_2, are equal and the optimal weighting factor of equation (11) is given by $\alpha = \frac{1}{2}$. The soft outputs, $y(n)$ of the diversity combining block can be expressed as,

$$y(n) = s(n) + \eta(n) \tag{12}$$

where $\eta(n)$ is the sum of the noise and the residual ISI components and,

$$\eta(n) = \frac{\eta_1(n) + \eta_2(n)}{2} \tag{13}$$

The combined channel-feedforward impulse response $\mathbf{h} = \mathbf{c} * \mathbf{f}$ is a L_h length vector, where $L_h = L_c + L_f - 1$. Let δ be the detection delay of the DFE. We define the vector g,

$$g(k) = \begin{cases} 0, & \delta \le k \le \delta + L_b \\ h(k), & \text{otherwise} \end{cases} \tag{14}$$

Recall that L_b is the length of the feedback filter. Equation (13) can be rewritten as,

$$\begin{aligned} \eta(n) = {} & \frac{1}{2}\sum_{k=0}^{L_h-1} g(k)\{s(n+\delta-k) + s(n-\delta+k)\} \\ & + \frac{1}{2}\sum_{k=0}^{L_f-1} f(k)\{w(n+\delta-k) + w(n+L_c-1-\delta+k)\} \end{aligned} \tag{15}$$

Equation (15) expresses the MSE of the overall system as a sum of the noise gain and the residual ISI terms.

$$\text{Noise gain term} = \sum_{k=k_{min}}^{k_{max}} \left\{ \frac{f(k) + f(2\delta + 1 - L_c - k)}{2} \right\}^2 \sigma_w^2 \tag{16}$$

where

$$k_{min} = \min(0, 2\delta + 1 - L_h) \quad \text{and} \quad k_{max} = \max(L_f - 1, 2\delta + 1 - L_c) \tag{17}$$

and

$$\text{Residual ISI term} = \sum_{k=j_{min}}^{j_{max}} \left\{ \frac{g(k) + g(2\delta - k)}{2} \right\}^2 \tag{18}$$

where

$$j_{min} = \min(0, 2\delta + 1 - L_h) \quad \text{and} \quad j_{max} = \max(L_h - 1, 2\delta) \tag{19}$$

For each individual stream, the noise gain term and the residual ISI terms are $\mathbf{f}^T\mathbf{f}\sigma_w^2$ and $\mathbf{g}^T\mathbf{g}$ respectively. From equations (16) and (18), we observe that the equal-gain diversity combining scheme has decreased[1] the contributions of both the noise gain and the residual ISI terms on the effective MSE when compared to either the normal-mode or time-reversal mode operation of the DFE. This diversity is a result of the non-linear structure of the DFE and the fact that we allow anti-causal processing of the received signal. For example, the output of a linear equalizer under time-reversal does not exhibit any diversity.

5 Optimal DFE for Equal-Gain Combining

Optimizing the MMSE-DFE taps to minimize the MSE at the decision device of each individual stream would not necessarily minimize the effective MSE of the receiver structure that incorporates diversity combining. In this section the optimization of the unbiased MMSE-DFE taps to minimize $\mathrm{E}[\eta^2]$, the effective MSE, for a symmetric channel is described.

The DFE taps are optimized for a fixed system delay, δ. The channel convolution matrix for the channel $\mathbf{c}$ is an $L_h \times L_f$ matrix and is,

$$C = \begin{bmatrix} c(0) & 0 & \cdots & 0 \\ \vdots & c(0) & \cdots & \vdots \\ c(L_c - 1) & \vdots & \cdots & 0 \\ 0 & c(L_c - 1) & \cdots & c(0) \\ \vdots & 0 & \cdots & \vdots \\ 0 & \vdots & \cdots & c(L_c - 1) \end{bmatrix} \tag{20}$$

---

[1] This can be easily proved using the Cauchy-Schwartz inequality

The combined channel feedforward impulse response can be expressed as $\mathbf{h} = C\mathbf{f}$. The optimal feedback filter taps of the DFE should exactly cancel the L_b taps of $\mathbf{h}$ that follow the cursor h_δ to minimize MSE. We define the $L_h \times L_h$ matrix V as,

$$V = \begin{bmatrix} \mathbf{I}_{\delta+1\times\delta+1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{L_b\times L_b} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix} \tag{21}$$

We define the vector $\mathbf{g}$ as

$$\mathbf{g} = VC\mathbf{f} - \mathbf{e}_\delta \tag{22}$$

where $\mathbf{e}_\delta$ is a column vector of length L_h and,

$$\mathbf{e}_\delta = [\mathbf{0}_{1\times\delta}\ 1\ \mathbf{0}_{1\times L_h-\delta-1}]^T \tag{23}$$

The variance of the residual ISI term in each of the streams is $\mathbf{g}^T\mathbf{g}$ and the combined residual ISI term of equation (18) can be expressed as

$$\text{Residual ISI term} = \frac{1}{4}(S_1 + S_2)^T\mathbf{g}^T\mathbf{g}(S_1 + S_2) \tag{24}$$

where S_1 and S_2 are $(2L_h - 1) \times L_h$ matrices that perform shifting and time-reversal operations on the vector $\mathbf{g}$, prior to averaging. The matrices S_1 and S_2 are constructed as,

$$S_1 = \begin{bmatrix} \mathbf{0}_{L_h-1-\delta\times L_h} \\ \mathbf{I}_{L_h\times L_h} \\ \mathbf{0}_{\delta\times L_h} \end{bmatrix} \quad ; \quad S_2 = \begin{bmatrix} \mathbf{0}_{\delta\times L_h} \\ \tilde{\mathbf{I}}_{L_h\times L_h} \\ \mathbf{0}_{L_h-1-\delta\times L_h} \end{bmatrix} \tag{25}$$

where $\mathbf{0}$ refers to a zero matrix, $\mathbf{I}$ refers to an identity matrix and $\tilde{\mathbf{I}}$ is a time-reversal matrix, i.e., with it's anti-diagonal elements as unity and all the other elements as zero. Similarly, the noise gain term of equation (16) can be expressed as,

$$\text{Noise gain term} = \frac{\sigma_w^2}{4}(M_1 + M_2)^T\mathbf{f}^T\mathbf{f}(M_1 + M_2) \tag{26}$$

where M_1 and M_2 are $(L_h + L_f - 1) \times L_f$ matrices that perform shifting and time-reversal operations on the vector $\mathbf{f}$ before averaging. The matrices M_1 and M_2 are constructed as shown below. If $2\delta + 1 - L_h \geq 0$ then,

$$M_1 = \begin{bmatrix} \mathbf{I}_{L_f\times L_f} \\ \mathbf{0}_{L_h-1\times L_f} \end{bmatrix} \quad ; \quad M_2 = \begin{bmatrix} \mathbf{0}_{2\delta+1-L_h\times L_f} \\ \tilde{\mathbf{I}}_{L_f\times L_f} \\ \mathbf{0}_{2L_h-2\delta-2\times L_f} \end{bmatrix} \tag{27}$$

and if $2\delta + 1 - L_h < 0$ then,

$$M_1 = \begin{bmatrix} \mathbf{0}_{L_h-2\delta-1\times L_f} \\ \mathbf{I}_{L_f\times L_f} \\ \mathbf{0}_{2\delta\times L_f} \end{bmatrix} \quad ; \quad M_2 = \begin{bmatrix} \tilde{\mathbf{I}}_{L_f\times L_f} \\ \mathbf{0}_{L_h-1\times L_f} \end{bmatrix} \tag{28}$$

The optimization of the DFE taps to minimize MSE is formulated as,

$$\min_{\mathbf{f}} \frac{1}{4}\Big[(S_1 + S_2)^T(VC\mathbf{f} - \mathbf{e}_\delta)^T(VC\mathbf{f} - \mathbf{e}_\delta)(S_1 + S_2) \\ + (M_1 + M_2)^T\mathbf{f}^T\mathbf{f}(M_1 + M_2)\sigma_w^2\Big] \tag{29}$$

The solution to the optimization problem of equation (29) is,

$$\mathbf{f} = [(VC)^T(S_1+S_2)^T(S_1+S_2)VC + \sigma_w^2(M_1+M_2)^T(M_1+M_2)]^{-1} \cdot (VC)^T(S_1+S_2)^T(S_1+S_2)\mathbf{e}_\delta \tag{30}$$

The unbiased DFE settings can be obtained from equation (30) by scaling $\mathbf{f}$ by the reciprocal of the cursor, $\frac{1}{h(\delta)}$.

$$\mathbf{f_{opt}} = \frac{\mathbf{f}}{h(\delta)} \tag{31}$$

and the optimum feedback filter is,

$$\mathbf{b_{opt}} = \frac{[h(\delta+1)\ \cdots\ h(\delta+L_b)]^T}{h(\delta)} \tag{32}$$

6 Simulation Results

An asymmetric baud-spaced channel impulse response of

$$\mathbf{c1} = [0.183\ \ 0.916\ \ 0.289\ \ -0.183\ \ 0.092\ \ -0.046\ \ 0.018]$$

was considered. An unbiased MMSE-DFE with 4 feedforward taps and 3 feedback taps was used for both the normal-mode operation and the time-reversal operation of the DFE. The DFE taps were optimized to minimize MSE at the decision device input for each stream. The MSE performance curves for a conventional DFE, time-reversal DFE and the proposed diversity combining scheme (see Figure 1) are compared in Figure 2. Correct decision feedback was assumed for computing the MSE values. The optimal weighting factor for the diversity combining scheme was computed using equation (11) and the MSE using equation (9). The suboptimal DFE refers to the scenario where the DFE taps were optimized to minimize MSE at the decision device input for each stream. The diversity combining scheme yields a gain of about 1 dB for an MSE of 0.1 (i.e., -10 dB).

Furthermore, the symbol error rate(SER) curves for the conventional DFE, time-reversal DFE and the diversity combining scheme are shown in Figure 2 for the asymmetric channel $\mathbf{c}_1$. The effect of error propagation is incorporated in these curves, i.e., correct decision feedback was not assumed. In the diversity combining scheme with hard DFE a slicer was used as the decision device. On the other hand, in the diversity combining scheme with soft DFE a soft decision device (see [8]) was used. The transmitted symbols were assumed to be from a BPSK constellation. The SER plots show gains similar to those of the MSE plots.

In another experiment a symmetric baud-spaced channel with an impulse response of

$$\mathbf{c_2} = [0.2887\ \ 0.9129\ \ 0.2887]$$

was considered. A DFE with 4 feedforward taps and 3 feedback taps was used and $\mathbf{c_2}$ being a symmetric channel, the DFE settings were the same for both the normal-mode operation and the time-reversal operation of the DFE. Figure 3 compares the MSE performance curves with and without the diversity combining scheme. The suboptimal DFE refers to the scenario where the DFE taps were optimized to minimize MSE at the decision device input for each stream.

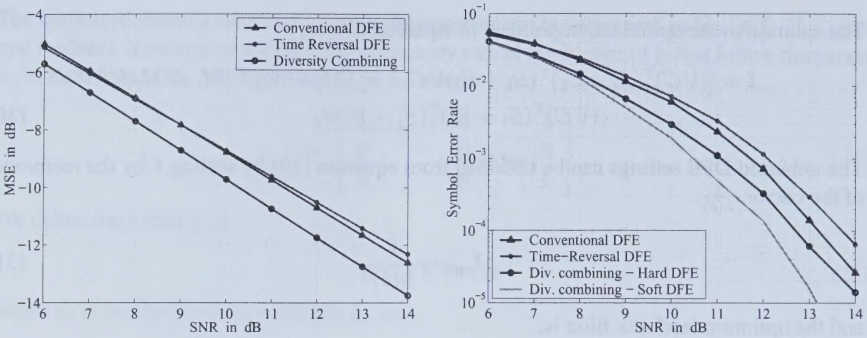


Figure 2: MSE and SER performance curves for an asymmetric channel

The optimal DFE, however, was optimized to minimize the MSE at the output of the diversity combining scheme and the DFE tap settings were computed using equations (31) and (32). Note that for the symmetric channel case the time-reversal mode operation will have the same performance as the normal-mode operation. Correct decision feedback was assumed while computing the MSE values.

The diversity combining scheme with the suboptimal DFE settings provides a gain of about 1 dB at an MSE of 0.1 over the normal-mode operation of the DFE. The diversity combining scheme with the optimal DFE tap settings provides an additional gain of 0.6 dB over the suboptimal case. However, in the presence of decision errors the same gains cannot be expected for the optimal DFE settings over the suboptimal DFE settings. This is due to the fact that the loss due to error propagation depends on the magnitude of the feedback filter coefficients and hence may be different for the two cases. Figure 3 also compares the SER plots for a conventional DFE, a DFE with ideal feedback and the diversity combining scheme with the suboptimal DFE tap settings. Gains similar to those of the MSE plots were observed.

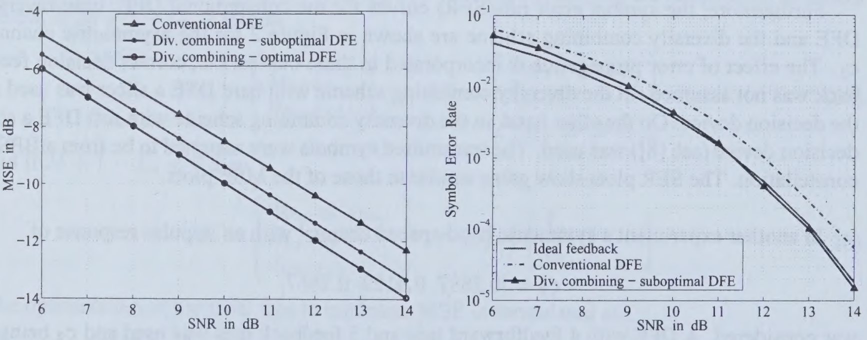


Figure 3: MSE and SER performance curves for a symmetric channel

7 Conclusions and Future Work

This paper considered the use of a time-reversal decision feedback equalizer (DFE) in conjunction with a conventional DFE. A receiver structure that combines the forward DFE outputs and the time-reversal DFE outputs to exploit the diversity between the two streams was proposed. An optimal combining scheme of the two streams that minimizes the mean-squared error (MSE) at the output of the receiver structure was derived. The inherent diversity between the soft outputs of the conventional DFE and a time-reversal DFE was quantified under a few simplifying assumptions. The DFE filter settings were optimized, for a channel with a symmetric impulse response, to minimize MSE at the output of the diversity combining block. Simulation results demonstrating performance gains for the proposed receiver structure were presented for both symmetric and asymmetric channels.

Current research focuses on the optimization of the DFE taps to minimize MSE at the output of the diversity combining block for an asymmetric channel. Furthermore, since the decisions at the output of the diversity combining block are more reliable than the tentative decisions in either of the normal-mode or the time-reversal DFE, it is possible to iteratively improve the performance by feeding back the more reliable decisions to the feedback filter of the DFE.

References

[1] M. Austin, "Decision feedback equalization for digital communication over dispersive channels," *MIT Research Laboratory of Electronics Technical Report 437*, August 1967.

[2] P. Monsen, "Feedback equalization for fading dispersive channels," *IEEE Transactions on Information Theory*, Vol.17, pp. 56-64, January 1971.

[3] H. Suzuki, "Performance of a new adaptive diversity-equalization for digital mobile radio," *Electron. Letters*, Vol. 26, No. 10, pp. 626-627, May 1990.

[4] S. Ariyavisitakul, "A decision-feedback equalizer with selective time-reversal operation for high-rate indoor radio communication," *GLOBECOM '90*, pp. 2035-2039, San Diego, CA, USA, December 1990.

[5] S. Ariyavisitakul, "A Decision Feedback Equalizer with Time-Reversal Structure," *IEEE Journal on Selected Areas in Communications*, Vol. 10, No. 3, pp. 599-613, April 1992.

[6] N. Al-Dhahir and J. M. Cioffi, "MMSE Decision-Feedback Equalizers: Finite-Length Results," *IEEE Transactions on Information Theory*, Vol. 41, No. 4, pp. 961-975, July 1995.

[7] C.S. McGahey, A.C. Singer and U. Madhow, "BAD: A Bi-directional Arbitrated Decision Feedback Equalizer," *Proc. of CISS 2000*, Vol. II, Princeton, NJ, USA, March 2000.

[8] J. Balakrishnan, H. Viswanathan and C.R. Johnson, Jr., "Decision Device Optimization for Soft Decision Feedback Equalization," *Proc. of CISS 2000*, Vol. II, Princeton, NJ, USA, March 2000.

[9] J.G. Proakis, "Digital Communications," *McGraw-Hill*, NewYork, 1995.

Application of Reduced-Rank Chip-level MMSE Equalization to the Forward Link DS-CDMA with Frequency Selective Multipath *

Samina Chowdhury, Michael D. Zoltowski and J. Scott Goldstein

School of Electrical Engineering,
Purdue University
West Lafayette, IN 47907-1285
email : {samina,mikedz}@ecn.purdue.edu

SAIC
4001 N. Fairfax Drive, Suite 400
Arlington, VA 22203
sgoldstein@trg1.saic.com

Abstract

This paper deals with synchronous Direct-Sequence CDMA transmission using orthogonal channel codes in frequency selective multi-path, motivated by the forward link in 3G CDMA systems. The chip-level MMSE estimate of the (multi-user) synchronous sum signal transmitted by the base, followed by a correlate and sum, has been shown to perform very well in saturated systems. In this paper, reduced-rank, chip-level MMSE estimation based on the Multi-stage Nested Wiener Filter (MSNWF) is presented.Our simulations show that, with perfect channel knowledge, only a small number of stages is needed to achieve near full-rank MSE performance over a practical SNR range. This is valid for the "edge of cell" scenario, where two base-stations are contributing equal-power signals, as well as the single base-station case. Also, adaptive MSNWF operating in a very low rank subspace and using the pilot channel for training is shown to perform slightly better than RLS and significantly better than LMS.

1 Introduction

Chip-level downlink equalizers have been proposed to significantly increase the capacity for high-speed wireless communication links, such as cdma2000. In this case, the multipath delay spread may span a significant portion of the symbol period, so that the multi-path propagation channel is frequency-selective. As a result the orthogonality of the Walsh-Hadamard spreading codes on the downlink is lost and there is significant multi-user access interference (MAI). When many or all users are active in the cell, the BER curve of the standard RAKE receiver flattens out at higher SNR [1]. Chip-rate MMSE equalizers that minimize the mean-square error between the synchronous sum signal of all active users transmitted from a given base station and its estimate were independently proposed by Ghauri et. al [2] and Krauss [1]. The MMSE estimators with perfect channel estimation were shown to outperform both ZF and RAKE [1].

However, for high data rate applications, the multi-path delay spread may span several chips and so the MMSE equalizer would require computation of a large number of

---

*This research was supported by the Air Force Office of Scientific Research under grant no. F49620-00-1-0127 and by Texas Instruments' DSP University Research Program.

coefficients, and may take an unacceptably long time to converge in adaptive implementations. In this paper, we present a reduced-rank, chip-level MMSE estimator based on the Multi-Stage Nested Wiener Filter(MSNWF) developed by Goldstein and Reed [3]. With perfect knowledge of the channel statistics, the MSNWF requires only a small number of stages to achieve near full-rank MMSE performance over a wide SNR range. We use a training-based block adaptive algorithm to simulate the performance of the MSNWF when the channel is unknown and the filter is adapted using a pilot channel and known pilot symbols. The SINR plot shows a convergence speed comparable to full-rank RLS and much faster than full-rank LMS. The superior performance of the MSNWF is further illustrated by simulated BER curves.

The results in this paper are for CDMA forward link with synchronous users, saturated loading, frequency selective fading, long code scrambling and employing two antennas at the mobile receiver. The channel is assumed to be unvarying with time, which might be valid only over a short time interval.

2 Data and Channel Model

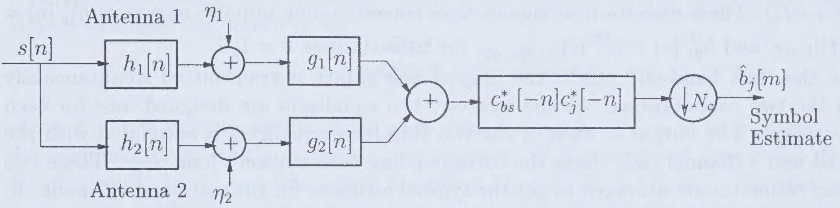


Figure 1: Chip-level MMSE Equalization for DS-CDMA downlink with 1 base-station.

The channel model is shown in Figure 1. If the transmission is from only one base-station, the impulse response for the i-th antenna channel between the transmitter and receiver (mobile station) is given by

$$h_i(t) = \sum_{k=0}^{N_m-1} h_i[k]p_{rc}(t-\tau_k) \qquad i = 1, 2. \tag{1}$$

where $p_{rc}(t)$ is the composite chip waveform (including the matched low-pass filters on the transmit and receive end). The chip waveform is assumed to have a raised cosine spectrum. N_m is the total number of delayed paths, i.e. 'multipath arrivals', some of which may have zero or negligible power, so that the channel impulse response is sparse.

The transmitted 'sum' signal may be described as

$$s[n] = c_{bs}[n] \sum_{j=1}^{N_u} \sum_{m=0}^{N_s-1} b_j[m]c_j[n - N_c m] \tag{2}$$

where $c_{bs}[n]$ is the base-station dependent long code, $b_j[m]$ is the bit/symbol sequence of the j-th user, $c_j[n]$ is the j-th user's channel (short) code of length N_c, N_u is the total number of active users, N_s is the number of bit/symbols transmitted during a given time window.

The signal received at the i-th antenna (after convolving with a matched filter having a square-root raised cosine impulse response) is given by

$$y_i(t) = \sum_n s[n]h_i(t - nT_c) + \eta_i(t) \qquad i = 1, 2 \tag{3}$$

where T_c is the duration of one chip and $\eta_i(t)$ is a noise process assumed white and Gaussian prior to coloration by the receiver chip-pulse matched filter .

2.1 Edge of Cell/Soft Handoff

We consider the interference problem when the desired user is at the edge of a cell so that the total received signal at the mobile station is the sum of the contributions from two base-stations, plus noise.

$$y_i(t) = y_i^{(1)}(t) + y_i^{(2)}(t) + \eta_i(t) \quad i = 1, 2 \text{ antennas at receiver.} \tag{4}$$

where the superscript denotes the corresponding base-station.

At each antenna, we oversample the signal $y_i(t)$ to obtain $y_{1i}[n] = y_i(nT_c)$ and $y_{2i}[n] = y_i(\frac{T_c}{2} + nT_c)$. These discrete-time signals have corresponding impulse responses $h_{1i}^{(k)}[n] = h_i^{(k)}(t)|_{t=nT_c}$ and $h_{2i}^{(k)}[n] = h_i^{(k)}(t)|_{t=\frac{T_c}{2}+nT_c}$ for base-stations $k = 1, 2$.

In the "soft hand-off" mode, the desired user's data is transmitted simultaneously from the two base-stations. At the receiver, two equalizers are designed, one for each base-station. The output of each of the two chip-level equalizers is correlated with the desired user's channel code times the corresponding base-station's long code. These two symbol estimates are averaged to get the symbol estimate for the soft hand-off mode. In the "normal mode", the second base-station is treated as interference.

3 Chip-level MMSE Estimator

The chip-level MMSE equalizer is designed to minimize the mean-squared error between the multi-user synchronous sum signal, $s[n]$ and the sum of the equalizer outputs, as depicted in Figure 1. Given the orthogonality of the channel codes, an estimate of the symbol, $\hat{b}_j[m]$ can be obtained via a correlate and sum with c_j and the base-station dependent long code at the output of the chip-level MMSE equalizer, once per symbol.

Equation (4) can be written more compactly in vector form as

$$\begin{aligned} \mathbf{y}[n] &= \mathbf{H}^{(1)}\mathbf{s}^{(1)}[n] + \mathbf{H}^{(2)}\mathbf{s}^{(2)}[n] + \boldsymbol{\eta}[n] \\ &= \boldsymbol{\mathcal{H}}\mathbf{s}[n] + \boldsymbol{\eta}[n] \end{aligned} \tag{5}$$

where $\mathbf{H}^{(l)}$ is the $2N_g \times (L + N_g - 1)$ channel convolution matrix, N_g is the length of the equalizer.

$$\boldsymbol{\mathcal{H}} = \begin{bmatrix} \mathbf{H}^{(1)} & \mathbf{H}^{(2)} \end{bmatrix} \qquad \mathbf{H}^{(l)} = \begin{bmatrix} \mathbf{H}_1^{(l)} \\ \mathbf{H}_2^{(l)} \end{bmatrix} \tag{6}$$

$$\mathbf{H}^{(l)} = \begin{bmatrix} h_i^{(l)}[0] & 0 & \ldots & 0 \\ h_i^{(l)}[1] & h_i^{(l)}[0] & 0 & 0 \\ \vdots & \ddots & \ddots & \ddots \\ h_i^{(l)}[L-1] & h_i^{(l)}[L-2] & \ddots & h_i^{(l)}[0] \\ 0 & h_i^{(l)}[L-1] & \ddots & h_i^{(l)}[1] \\ \vdots & \ddots & \ddots & \ddots \\ 0 & 0 & 0 & h_i^{(l)}[L-1] \end{bmatrix}^T . \tag{7}$$

The corresponding formulae for only one transmitting base-station can be found by simply removing $\mathbf{H}^{(2)}$.

Krauss and Zoltowski [1] made some simplifying assumptions to derive a chip-level MMSE equalizer that can be easily implemented. The sequence values for the multi-user sum signal are assumed to be i.i.d. random variables. Otherwise the covariance matrix of the sum signal $\mathbf{s}[n]$ is a complicated expression involving the Walsh-Hadamard spreading codes that vary from index to index. The i.i.d. assumption is valid if the (long) scrambling code is viewed as a random i.i.d. sequence and/or all users are active with equal power. With this assumption, the covariance matrix of the signal is $E\{\mathbf{s}[n]^H \mathbf{s}[n]\} = \sigma_s^2 \mathbf{I}$ and the MMSE equalizer was shown to be

$$\mathbf{g}^{(l)} = \{\sigma_s^2 \boldsymbol{\mathcal{H}}\boldsymbol{\mathcal{H}}^H + R_{\eta\eta}\}^{-1} \mathbf{H}^{(l)} \boldsymbol{\delta}_{D_c} \tag{8}$$

where $\boldsymbol{\delta}_{D_c}$ is a column vector of all zeros except 1 in the $(D_c + 1)$-th position, D_c is the combined delay of the equalizer and channel, σ_s^2 is the signal power and $R_{\eta\eta} = E\left[\boldsymbol{\eta}^H[n]\boldsymbol{\eta}[n]\right]$. In [1] and this paper, the delay D_c, $0 \leq D_c \leq (N_g + L - 2)$ that yields the smallest MMSE is calculated using the actual channel statistics and that D_c is used in the simulations.

Equation (8) has the form of the well-known Wiener-Hopf solution

$$\mathbf{w} = \mathbf{R}_{xx}^{-1} \mathbf{r}_{dx} \tag{9}$$

where $\mathbf{R}_{xx}$ is the channel covariance matrix and $\mathbf{r}_{dx}$ is the cross-correlation vector.

In [1], [4], Krauss and Zoltowski showed that the MMSE significantly outperforms the RAKE receiver. This is clearly illustrated in Figure 2, which plots the difference in SNR between the RAKE and MMSE receivers as a function of target uncoded (theoretical) BER, for reception at the edge of a cell. We refer to [4] for details of the simulation results. For both normal and soft hand-off modes of operation, the RAKE requires much more power than the MMSE receiver. This is more pronounced when soft hand-off is unavailable. So MMSE equalization would allow operation in SNR regions that would be impossible with RAKE receivers - especially when a large number of channel codes are active relative to the spreading factor.

4 Reduced Rank Filtering

Recently, reduced rank techniques have received attention in the context of adaptive linear equalization for DS-CDMA systems. The received signal vector is projected onto a lower dimensional subspace, and the Wiener filter given by Equation (9) is constrained to lie in this subspace. This has the advantage of reducing the number of filter coefficients

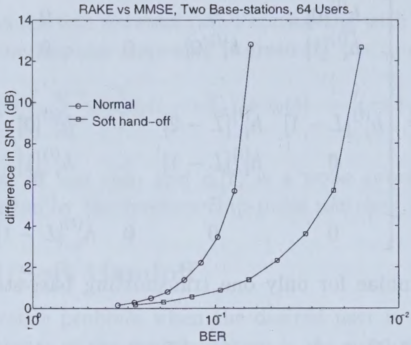


Figure 2: Improvement in SNR from RAKE to MMSE, Two Base-stations

to be estimated, and so increases the speed of convergence dramatically for adaptive methods, if the subspace is chosen properly. But the overall MMSE for the reduced-rank filter may be higher than the MMSE for the full-rank filter. The most widely known reduced rank techniques are the Principal Components method [5] and the Cross Spectral methods [6], both based on eigen-decomposition of the channel covariance matrix, $\mathbf{R}_{xx}$.

4.1 Multi-Stage Nested Wiener Filter

Goldstein and Reed [3] first formulated the MSNWF, which uses the information from the channel covariance matrix, $\mathbf{R}_{xx}$ and cross-correlation vector, $\mathbf{r}_{dx}$ to determine the bases of the lower-dimension subpace. The structure of the MSNWF is depicted in Figure 3. The observed vector process is decomposed by a sequence of nested filters $\mathbf{p}_1, \mathbf{p}_2, \cdots, \mathbf{p}_D$, where D is the order of the filter, and

$$\mathbf{p}_k = \frac{E[\mathbf{x}_{k-1}\, d_{k-1}^*]}{||E[\mathbf{x}_{k-1}\, d_{k-1}^*]||} \tag{10}$$

The input process to the $(k+1)$-th stage is $\mathbf{x}_k[n] = \mathbf{B}_k \mathbf{x}_{k-1}[n]$, where $\mathbf{B}_k$ is a 'blocking' matrix such that $\mathbf{B}_k \mathbf{p}_k = 0$.

The outputs of the various stages are linearly combined via the *scalar* weights $w_1, w_2, ..,$ w_{D-1}, chosen so that the mean-square error at each stage is minimized. If the decomposition is carried out for the full N stages, then the multi-stage nested filter is exactly equivalent to the full-rank classical Wiener filter[3]. Furthermore, it is optimal in terms of reducing the MSE for a given rank.

Honig and Xiao [7] first proposed choosing a projection matrix onto the subspace orthogonal to $\mathbf{p}_k$ as the blocking matrix at each stage, hence retaining the length N of the filter and 'observed signal' $\mathbf{x}_k[n]$. With this choice for the blocking matrix, $\mathbf{T}_D = [\mathbf{p}_1\ \mathbf{p}_2\ ...\ \mathbf{p}_D]$ forms an orthonormal basis for the reduced dimension subspace. Moreover, $\mathbf{w}$ is constrained to lie in the Krylov subspace spanned by $\{\mathbf{r}_{dx}, \mathbf{R}_{xx}\mathbf{r}_{dx}, \mathbf{R}_{xx}^2\mathbf{r}_{dx}, \ldots \mathbf{R}_{xx}^{D-1}\mathbf{r}_{dx}\}$ [7]. The orthogonal decomposition ensures that the reduced dimension $D \times D$ correlation matrix $\mathbf{T}_D^H \mathbf{R}_{xx} \mathbf{T}_D$ is tri-diagonal [3].

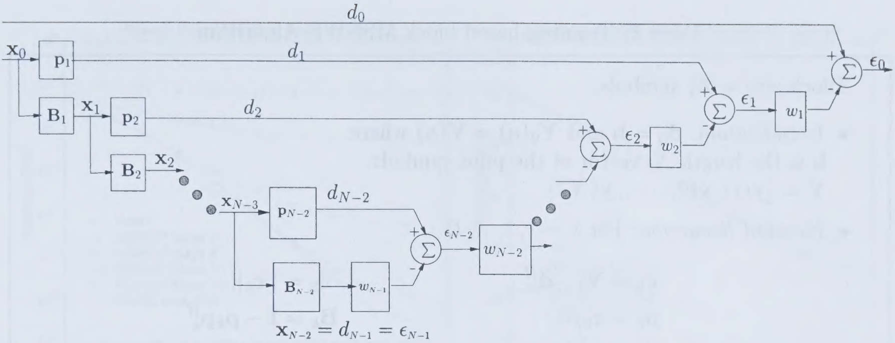


Figure 3: Structure of successive stages of the Multistage Nested Weiner Filter.

4.2 Application of MSNWF to CDMA Downlink

Our first set of results solves for chip-level MMSE equalization based on Equation (8) when only one base-station is transmitting and finds the "ideal" MSNWF solution after various stages, assuming $\mathbf{R}_{xx}$ and $\mathbf{r}_{dx}$ are known (perfect channel estimation). The same simulations are also done for the edge of cell situation, when two base-stations of equal power are received at the mobile-station and the receiver uses soft handoff.

Next we use the class of training-based adaptive algorithms presented by Honig and Goldstein in [8] to simulate the performance of MSNWF when the channel is unknown. The algorithm is given in Table 1. Although the MMSE equalizer described in this paper estimates the chip-rate multi-user synchronous sum signal, it is not possible to train the equalizer on this signal as that would require the knowledge of number of active users, all of the active channel codes and the transmitted symbols. Instead, we use the pilot channel of CDMA downlink, which has a known code and known symbols. For DS-CDMA with orthogonal spreading codes, the MMSE equalizer is identical for all channels within a multiplicative constant, so the pilot code can be used to train for any other channel code [9]. The received chip-rate signal is arranged into length N_c vectors, correlated with the pilot channel code times the appropriate portion of the base-station long code and then filtered by the equalizer. The output is the 'pilot symbol estimate'.

5 Simulation Results

A wideband CDMA forward link was simulated similar to one of the options in the US cdma2000 proposal. The chip rate was 3.6864 MHz ($T_c = 0.27\mu s$), 3 times that of IS-95. Simulations were performed for a "saturated cell", i.e. all 64 possible channel codes are active, and all users have *equal* power. The spreading factor was $N_c = 64$ chips per bit. The data symbols were BPSK, and spread with one of 64 Walsh-Hadamard codes with length 64 for each user. All users were of equal power, and their signals were summed synchronously and then multiplied with a QPSK scrambling code of length 32678, similar to the IS-95 standard.

The channels were modeled to have four equal-power multi-paths, the first one arriving at 0, the last at $10\mu s$ (corresponding to about 37 chips) and the other two delays picked at random in between. The multi-path coefficients are complex normal, independent

Table 1: Training-based block MSNWF Algorithm

Block size = N_t symbols.

- *Initialization:* $\mathbf{d}_0 = \mathbf{b}$ and $\mathbf{Y}_0(n) = \mathbf{Y}(n)$ where $\mathbf{b}$ is the length N_t vector of the pilot symbols, $\mathbf{Y} = [\mathbf{y}(1), \mathbf{y}(2), \cdots, \mathbf{y}(N_t)]$.
- *Forward Recursion:* For $k = 1, 2, ..., D$:

$$\begin{aligned} \mathbf{c}_k &= \mathbf{Y}_{k-1}\mathbf{d}_{k-1}^H & \delta_k &= \|\mathbf{c}_k\| \\ \mathbf{p}_k &= \mathbf{c}_k/\delta_k & \mathbf{B}_k &= \mathbf{I} - \mathbf{p}_k\mathbf{p}_k^H \\ \mathbf{d}_k &= \mathbf{p}_k^H\mathbf{Y}_{k-1} & \mathbf{Y}_k &= \mathbf{B}_k\mathbf{Y}_{k-1} \end{aligned}$$

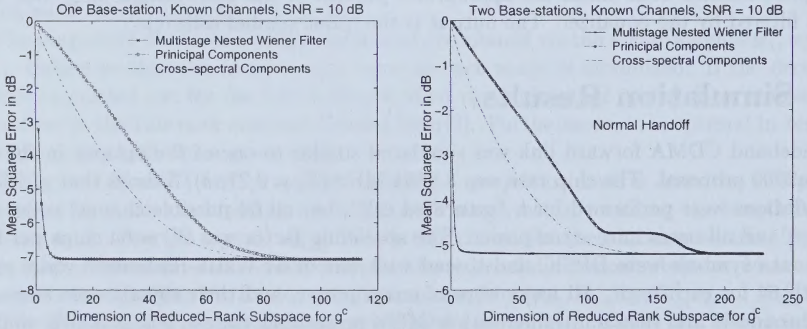
- *Backward Recursion:* For $k = D, D-1, ..., 1$, with $\epsilon_D(n) = d_D(n)$:

$$\begin{aligned} w_k &= (\epsilon_k\mathbf{d}_{k-1}^H)/\|\epsilon_k\|^2 = \delta_k/\|\epsilon_k\|^2 \\ \epsilon_{k-1}(n) &= \mathbf{d}_{n-1}(n) - w_k^*\epsilon_k(n) \end{aligned}$$

random variables with equal variance. The arrival times at antenna 1 and 2 are the same, but the multi-path coefficients are independent.

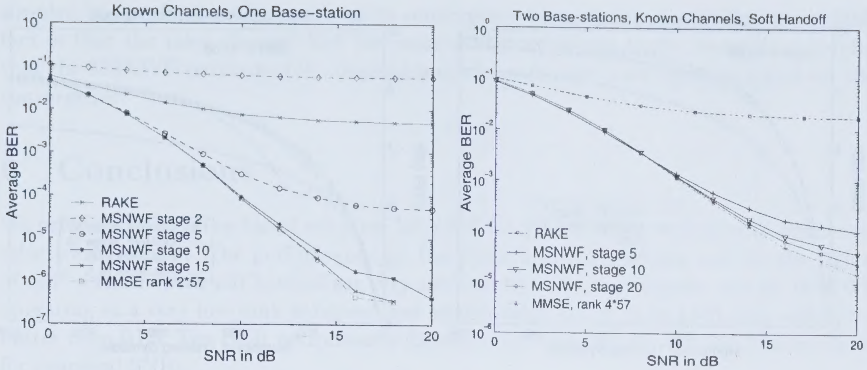
In the two base-station case, the channels are scaled so that the total energy from each of the two base-stations is equal at the receiver. The 4 multi-path arrivals from the 2nd base-stations are random, with maximum delay spread of $10\mu s$. SNR is defined to be the ratio of the sum of the average power of the received signals over all the channels, to the average noise power, after chip-matched filtering. The curves were generated by averaging over many different channels. Note that the abscissa in the graphs is the **post**-correlation SNR for *each* user which includes a processing gain of 10log(64) $\approx$ 18 dB.

Figure 4 plots the Mean-Square Error for the different reduced-rank methods as a



(a) (b)

Figure 4: MSE vs Rank of Reduced Dimension Subspace



(a)

(b)

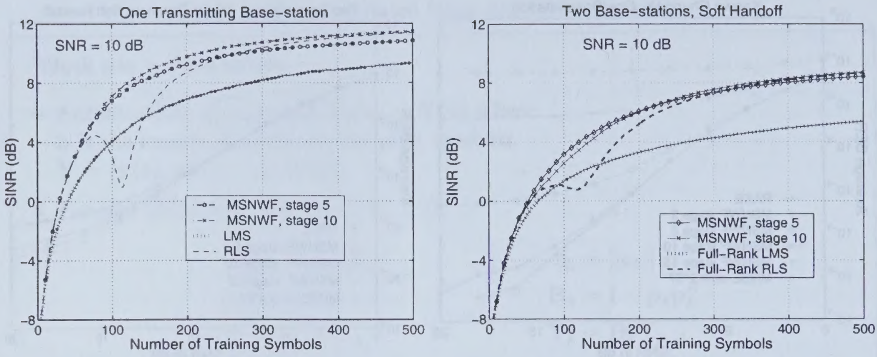
Figure 5: BER for Different Chip-level Equalizers for CDMA Downlink.

function of the subspace dimension, D. The channel statistics and noise power are assumed to be known (i.e. perfect channel estimation). In the single base-station case,4(a), the dimension of the full space is 114 (the equalizer length is 57 at each of 2 antennas, as multipath delay spread is 37 chips and the chip pulse waveform is cut off after 5 chips at both ends). The MSE for MSNWF is seen to drop dramatically with D, and achieves the performance of the full-rank Wiener filter at dimension approximately 7! In contrast, the Principal components method takes longer than twice the delay spread, and the Cross-spectral method does only slightly better.

Figure 5(a) displays the BER curves obtained with the MSNWF for different sizes of the reduced-dimension subspace. The channel statistics are assumed to be known perfectly, so these curves serve as an informative upper bound on the performance. It is observed that even a 2-stage reduced-rank filter outperforms the RAKE at all SNR's and only a small number of stages of the MSNWF are needed in order to achieve near full-rank MMSE performance over a practical range of SNR's.

Figures 4(b) and 5(b) display the similar plots, but for the "edge of cell" scenario. In this case, there are 4 effective channels at the receiver, because we sampled the received signal at twice the chip-rate at each antenna. The dimension of the full space is 228 which makes the full rank calculations very cumbersome. Amazingly, the MSE for MSNWF still goes down very steeply with rank and achieves the full-rank value for subspace dimension of only 8 or so. In the BER plots of Figure 5(b), the bit error is calculated for the "soft handoff" mode. With perfect channel estimation, the MSNWF can achieve error rates similar to the full-rank MMSE over practical SNR range after stopping at stage as low as 5!

These plots suggest that MSNWF can achieve rapid adaptation in the case where the chip-level MMSE equalizer is adapted based on a pilot channel. Figure 6 plots the output SINR for different chip-level equalizers vs. time in symbols, at a fixed SNR. The output SINR is calculated using the formula derived by Krauss in [4]. The MSNWF, at stages 5 and 10 show very good performance with low sample-support. The convergence rate is similar to that of a full-rank RLS, which even asymptotically, does not beat the MSNWF



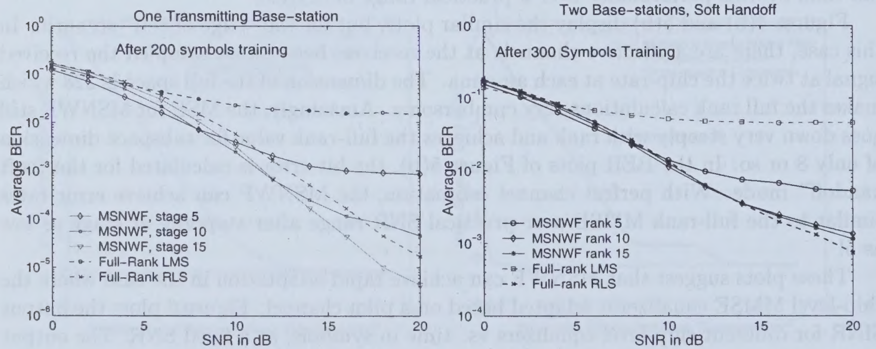
(a) (b)

Figure 6: Output SINR vs Time for Adaptive Chip-level Equalizers for CDMA downlink.

of rank only 10! The LMS algorithm converges much slower and to a lower SINR. For the two base-station case, the asympotic SINR is lower for all the equalizers due to the added interference from the MAI of the 2nd base-station. But the convergence speed of low-rank MSNWF is still impressive.

The BER curves in Figure 7 illustrate the performance of these equalizers. At low SNR's, the BER for MSNWF stage 5 is actually slightly lower than the BER for stage 10 or 15. This can be explained as the 'penalty' for learning the channels, i.e. in the presence of significant noise, the MSNWF with the fewer coefficients to adjust performs better. But as the SNR is increased, the signal power is sufficient for training, and the higher stages of MSNWF yield better approximation of the full-rank filter.

It is noteworthy that over a practical SNR range, in this adaptive implementation, the stage 5 MSNWF does better or almost as good as full-rank RLS! This improvement comes with much lower computational complexity than the RLS. The LMS algorithm is



(a) (b)

Figure 7: BER for Adaptive Chip-level Equalizers for CDMA Downlink.

simpler, but performs quite poorly as its convergence is much slower. Another remarkable fact is that the pilot channel had the same power as all the traffic channels, implying that the MSNWF reduced-rank adaptation does not require a strong pilot signal for fast convergence.

6 Conclusions

We presented a training-based adaptive MSNWF for the CDMA downlink with frequency-selective multipath. The performance for the single base-station case, and for the "edge-of-cell" scenario with soft handoff are very satisfactory. The convergence rate for MSNWF operating in a very low-rank subspace was significantly better than LMS, and somewhat better than RLS. The BER performance showed improvement over the full-rank methods for practical SNRs.

References

[1] M. D. Zoltowski T. P. Krauss and G. Leus. Simple MMSE Equalizers for CDMA Downlink to Restore Chip Sequence: Comparison to Zero-Forcing and RAKE. In *Intl. Conf. on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, 5-9 June 2000.

[2] I. Ghauri and D. T. M. Slock. Linear Receivers for the DS-CDMA Downlink Exploiting Orthogonality of Spreading Sequences. In *Proc. 32nd Asilomar Conf. on Signals, Systems, and Computers*, Pacific Grove, CA, November 1998.

[3] I. S. Reed J. S. Goldstein and L. L. Scharf. A Multistage Representation of the Wiener Filter based on Orthogonal Projections. *IEEE Trans. Information Theory*, 44(7):2943–2959, Nov. 1998.

[4] T. P. Krauss and M. D. Zoltowski. MMSE Equalization Under Conditions of Soft Hand-Off. In *IEEE Sixth International Symposium on Spread Spectrum Techniques & Applications (ISSSTA 2000)*, NJIT, New Jersey, 6-8 Sept. 2000.

[5] X. Wang and H. V. Poor. Blind Multiuser Detection. *IEEE Trans. on Information Theory*, 44(2):677–690, March 1998.

[6] J. S. Goldstein and I. S. Reed. Reduced-Rank Adaptive Filtering. *IEEE Trans. on Signal Processing*, 43(2):492–496, February 1997.

[7] M. L. Honig and W. Xiao. Large System Performance of Reduced-rank Linear Interference Supression for DS-CDMA. In *Proc. Allerton Conf. on Comm.,Control and Computing*, UIUC, October 1999.

[8] M. L. Honig and J. S. Goldstein. Adaptive Reduced-Rank Residual Correlation Algorithms for DS-CDMA Interference Suppression. In *Proc. 32ndAsilomar Conf. on Signals,Systems and Computing*, Pacific Grove, CA, November 1998.

[9] Colin D. Frank and Eugene Visotsky. Adaptive Interference Suppression for Direct-Sequence CDMA Systems with Long Spreading Codes. In *Proceedings 36th Allerton Conf. on Communication, Control, and Computing*, pages 411–420, Monticello, IL, 23-25 Sept. 1998.

“Turbo equalization” using frequency domain equalizers

Michael Tüchler, Joachim Hagenauer
Institute for Communications Engineering
Munich University of Technology
80290 München, Germany
micha@lnt.ei.tum.de, hag@lnt.ei.tum.de

Abstract

For coded data transmission over channels introducing inter-symbol interference (ISI), Douillard et al. proposed the “turbo equalization” approach, which is an iterative equalization and decoding algorithm for the receiver able to almost entirely compensate for the system performance degradation, e.g., in the bit error rate, due to ISI. One drawback of this approach is the exponentially increasing (in the ISI channel response length) complexity of the equalization step. To overcome this, several equalizers applicable to “turbo equalization” requiring reduced computational complexity have been proposed. We rederive some of these approaches and equalize the received data in the frequency domain. Besides the promising performance results, the new algorithms provide a further complexity reduction compared to the time domain counterparts.

1 Introduction

Many communication systems today encounter the problem of data transmission over a channel with inter-symbol interference (ISI). To protect the integrity of the data to be transmitted, a controlled amount of redundancy is added (encoding) using an error correction code (ECC). In this paper, we assume a coherent symbol-spaced receiver front-end, and that the receiver knows the signal phase and symbol timing precisely, such that the channel can be approximated by an equivalent, discrete-time, baseband model, as shown in Figure 1, where the transmit filter, the channel, and the receive filter, are represented by a discrete-time *linear* filter, with the finite-length impulse response (FIR)

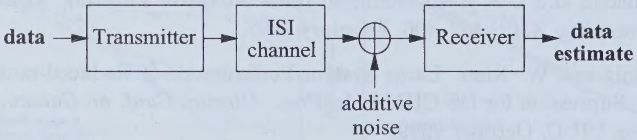


Figure 1: Discrete-time baseband representation of the data transmission system.

$$h[n] = \sum_{k=0}^{M-1} h_k \, \delta[n-k] \tag{1}$$

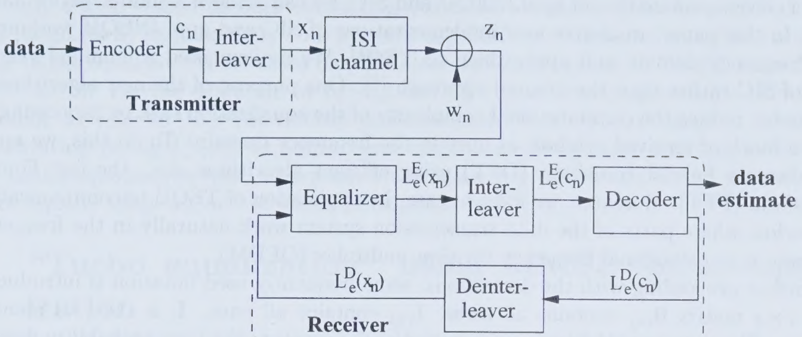


Figure 2: Data transmission system with a receiver performing "turbo equalization".

approach	complexity
MAP equalization	$O(q^M)$
SIC	$O(M)$
appLINEQU	$O(N+M)$
LINEQU	$O(N^2+M^2)$

Table 1: Computational complexity of equalization algorithms for "turbo equalization".

of length M. The coefficients h_k are assumed to be time-invariant and known to the receiver. The noise process is assumed to be independent and identically distributed (i.i.d.) and independent of the data.

Out of the various ways to implement such a transmission system, we look at the "turbo equalization" (TEQU) scheme, depicted in Figure 2, as one possible strategy for joint equalization and decoding. TEQU is an iterative receiver algorithm for coded data transmission applying the "turbo principle", where equalization and decoding tasks are repeated on the same set of received data while feedback information from the decoder is incorporated into the equalization process The "turbo principle" [1, 2] was originally developed for concatenated convolutional codes ("turbo coding" [3]) and is now adapted to various communication problems such as trellis coded modulation (TCM) and code division multiple access (CDMA). TEQU systems were proposed in several articles [4, 5, 6, 7, 8]. They differ in the chosen algorithm for equalization, e.g., maximum a-posteriori probability symbol estimation (MAP) [5, 6], maximum-likelihood sequence estimation (ML) [4, 5], or an approach using linear filtering [7, 8]. In the decoder, all of them use exclusively ML/MAP-based techniques applied to convolutional codes. The major motivation for developing these different TEQU algorithms was to decrease the computational complexity of the equalization task, which we describe with the order $O(\cdot)$ [9]. Table 1 shows this order per transmitted symbol (using a size q signal alphabet) and iteration for MAP equalization (MAPEQU) [5], soft interference cancellation (SIC) [7], and linear equalization (LINEQU) with a length N FIR filter employing the minimum mean squared error (MMSE) constraint [8]. The latter approach can be implemented in an exact and approximate way denoted as LINEQU and appLINEQU, respectively.

It was shown [8] that a pure MAP-based solution for equalization and decoding is fastest in the bit error rate (BER) improvement over successive iterations and achieves these improvements at the lowest signal-to-noise ratios (SNRs) compared to all other approaches. Unfortunately, for long channels (large M) or high order signal constellations (large q), MAPEQU quickly becomes unfeasible (see Table 1). However, using LINEQU and a clever combination of appLINEQU and SIC, we can achieve a similar performance [10]. In this paper, we derive new implementations of SIC and appLINEQU working in the frequency domain and apply them to TEQU. We implemented a modified version [10] of SIC rather than the original approach [7]. One purpose of the new algorithms is to further reduce the computational complexity of the equalization task by processing an entire block of received symbols at once in the frequency domain. To do this, we apply the discrete Fourier transform (DFT) using efficient algorithms, e.g., the fast Fourier transform (FFT). Moreover, we want to ease the application of TEQU to communication scenarios, where parts of the data transmission system work naturally in the frequency domain, e.g., orthogonal frequency division multiplex (OFDM).

Before proceeding with the derivations, some frequently used notation is introduced. The $i \times j$ matrix $\mathbf{0}_{i\times j}$ contains all zeros, $\mathbf{1}_{i\times j}$ contains all ones. $\mathbf{I}_i$ is the $i\times i$ identity matrix. The operator $E\{\cdot\}$ is the expectation with respect to the joint probability density function (pdf) of the x_n and w_n. The covariance operator $\mathrm{Cov}(\mathbf{x}, \mathbf{y})$ equals $E\{\mathbf{x}\mathbf{y}^H\} - E\{\mathbf{x}\}E\{\mathbf{y}^H\}$, where ^{H} is the hermitian operator. The operator $\mathbf{Diag}\,[x_0 x_1 \cdots x_{N-1}]$ applied to a length N vector returns a $N \times N$ square matrix with the vector elements along the diagonal. The operator $\mathbf{Circ_K}[x_0 x_1 \cdots x_{N-1}]$ applied to a length N vector $(K \geq N)$ returns the $K \times K$ circulant matrix

$$\mathbf{Circ_K}[x_0 x_1 \cdots x_{N-1}] \triangleq \begin{bmatrix} x_0 & 0 & \cdots & & 0 & x_{N-1} & x_{N-2} & \cdots & x_1 \\ x_1 & x_0 & 0 & \cdots & & 0 & x_{N-1} & \cdots & x_2 \\ & & & \ddots & & & \ddots & & \\ 0 & \cdots & & 0 & x_{N-1} & x_{N-2} & \cdots & & x_0 \end{bmatrix}. \tag{2}$$

To simplify the derivations, we define a specific data transmission system according to Figure 2. Data is encoded blockwise with a (binary) convolutional encoder yielding L code symbols $c_n \in \{+1, -1\}$, $n = 0, 1, ..., L-1$, already mapped to a signal constellation alphabet for binary phase shift keying (BPSK). We assume that $L \geq M$. The interleaver $\Pi(\cdot)$, a fixed random permutation on L elements, permutes the sequence $\mathbf{c} \triangleq [c_0\, c_1 \cdots c_{L-1}]^T$ and outputs $\mathbf{x} \triangleq [x_0\, x_1 \cdots x_{L-1}]^T = \Pi(\mathbf{c})$. The permutation function $\Pi^{-1}(\cdot)$, the deinterleaver, reverses the $\Pi(\cdot)$ operation. Transmitted over the ISI channel are the last $M-1$ symbols of the block $\mathbf{x}$ followed by $\mathbf{x}$ itself:

$$\tilde{\mathbf{x}} = [\, x_{L-M+1}\, x_{L-M+2} \cdots x_{L-1}\; x_0\, x_1 \cdots x_{L-1}\,]^T. \tag{3}$$

This augmentation, the added sequence is also called guard interval in OFDM, does not necessarily result in a rate loss, since many block-based transmission systems already match this requirement due do header and/or tail sequences at the block ends. By applying this step, the block $\mathbf{z} \triangleq [z_0\, z_1 \cdots z_{L-1}]^T$ of received symbols z_n is given as

$$\begin{aligned} \mathbf{H}_C &\triangleq \mathbf{Circ_L}[h_0\, h_1 \cdots h_{M-1}], \\ \mathbf{z} &= \mathbf{H}_C\, [x_0\, x_1 \cdots x_{L-1}]^T + [w_0\, w_1 \cdots w_{L-1}]^T, \end{aligned} \tag{4}$$

using Eq. (1) and Figure 2. The first $M-1$ symbols z_n were discarded, since they correspond to the transmitted guard interval. The noise is modeled as additive white Gaussian noise (AWGN), i.e., the noise samples w_n are i.i.d. with normal pdf $f_w(w) = \frac{1}{\sqrt{2\pi}\,\sigma_w}\exp(-\frac{w^2}{2\sigma_w^2})$. The equalizer processes the block $\mathbf{z}$ and outputs the log-likelihood ratios (LLRs) [1] $L_e^E(x_n)$, $n = 0, 1, ..., L-1$, about the transmitted symbols x_n. After deinterleaving, the LLRs $L_e^E(x_n)$ reordered to $L_e^E(c_n)$ are input to the decoder. We will use only a MAP-based decoder outputting quantized estimates of the transmitted data and the LLRs $L_e^D(c_n)$ about the code symbols c_n, which are interleaved to $L_e^D(x_n)$ and fed back to the equalizer. MAP-based convolutional decoding applied to TEQU is described in [1, 5, 10]. For the definition of $L_e^E(x_n)$ and $L_e^D(c_n)$ and the requirements they have to fulfill, we refer to [10]. The receiver processes each block $\mathbf{z}$ several times by repeating equalization (using $L_e^D(x_n)$ as a-priori information on the x_n) and decoding tasks. For the initial equalization task, no a-priori information is available and the x_n are assumed to be equally likely $+1$ or -1, which corresponds to $L_e^D(x_n)=0$, $\forall n$.

2 "Turbo equalization" using MMSE linear equalization

For this approach derived in [12] and applied to TEQU in [8], a linear equalizer consisting of a FIR filter with coefficients $c_{k,n}$, $k=-N_1, -N_1+1, \cdots, N_2$, of length $N=N_1+N_2+1$ is used as MMSE estimator outputting the estimates $\hat{x}_n$ of the transmitted symbols x_n. The design rule to obtain the $c_{k,n}$ is to minimize the MMSE cost function $E\{|x_n-\hat{x}_n|^2\}$. Assume that $N_1=L-1$ and $N_2=0$. In this case, the estimation is based on L received symbols z_n combined in vector $\mathbf{z}_n \triangleq [z_n\, z_{n+1} \cdots z_{n+L-1}]^T$. Given the coefficent vector $\mathbf{c}_n \triangleq [c_{0,n}^*\, c_{-1,n}^* \cdots c_{-L+1,n}^*]^T$, the estimates $\hat{x}_n$ are computed as follows [12]:

$$\begin{aligned} \mathbf{c}_n &= \mathrm{Cov}(\mathbf{z}_n, \mathbf{z}_n)^{-1}\mathrm{Cov}(\mathbf{z}_n, x_n), \\ \hat{x}_n &= E\{x_n\} + \mathbf{c}_n^H(\mathbf{z}_n - E\{\mathbf{z}_n\}). \end{aligned} \tag{5}$$

The quantities $\mathrm{Cov}(\mathbf{z}_n, \mathbf{z}_n)$ and $\mathrm{Cov}(\mathbf{z}_n, x_n)$ can be expressed in terms of $E\{x_n\}$ and $\mathrm{Cov}(x_n, x_m)$, which both depend on the a-priori information $L_e^D(x_n)$ [12]:

$$\begin{aligned} &E\{x_n\} = \sum_{k\in\{+1,-1\}} \Pr\{x_n=k\}\cdot k = \tanh\left(\frac{1}{2}\ln\frac{\Pr\{x_n=+1\}}{\Pr\{x_n=-1\}}\right) = \tanh\left(\frac{L_e^D(x_n)}{2}\right), \\ &\mathrm{Cov}(x_n, x_m) = 1 - E\{x_n\}E\{x_m\}^*. \end{aligned} \tag{6}$$

Using the approximate implementation of this algorithm, the coefficients $\mathbf{c}_n$ are computed at $n=0$ under the assumption that the x_n are equally likely $+1$ or -1 yielding $E\{x_n\}=0$ and $\mathrm{Cov}(x_n, x_m)=\delta[n-m]$ [12]. In this case, we have $\mathbf{z}_0=\mathbf{z}$ using Eq. (4) and therefore

$$\begin{aligned} \mathrm{Cov}(\mathbf{z}_0, \mathbf{z}_0) &= \sigma_w^2\mathbf{I}_L + \mathbf{H}_C\mathbf{H}_C^H, \\ \mathrm{Cov}(\mathbf{z}_0, x_0) &= \mathbf{H}_C\,[\,1\ \mathbf{0}_{1\times(L-1)}\,]^T. \end{aligned}$$

The time-invariant coefficient vector $\mathbf{c}$ is then

$$\begin{aligned} \mathbf{u} &\triangleq [\,1\ \mathbf{0}_{1\times(L-1)}\,]^T, \\ \mathbf{c} &\triangleq \mathbf{c}_0 = (\sigma_w^2\mathbf{I}_L + \mathbf{H}_C\mathbf{H}_C^H)^{-1}\,\mathbf{H}_C\mathbf{u}, \end{aligned} \tag{7}$$

which is used to equalize the entire block $\mathbf{z}$. With Eq. (5), the estimates $\hat{x}_n$ are computed by filtering $z_k-E\{z_k\}$, $k=n, n+1, ..., n+L-1$, with $\mathbf{c}$ and adding $E\{x_n\}$. Since $\hat{x}_n$ must not depend on $L_e^D(x_n)$ ("extrinsic" information only), the influence of $L_e^D(x_n)$ on $\hat{x}_n$ is subtracted off [12]. This yields altogether

$$\begin{aligned} \boldsymbol{\lambda}_n &\triangleq [\,E\{x_{n-L+1}\}\,E\{x_{n-L+2}\}\cdots E\{x_n\}\,]^T = [\,\lambda_{n-L+1}\,\lambda_{n-L+2}\cdots\lambda_n\,]^T, \\ \hat{x}_n &= \mathbf{c}^H\,(\mathbf{z}_n - \mathbf{H}_C\,\boldsymbol{\lambda}_n + \lambda_n\,\mathbf{H}_C\mathbf{u}), \end{aligned} \tag{8}$$

where we call the mean of x_n, $\lambda_n \triangleq E\{x_n\}$, soft bit. Using the non-causal length L linear filter $\mathbf{c}$, we need symbols z_m, $n \leq m \leq (L-1+n)$ outside the boundaries of block $\mathbf{z}$ for $n > 0$. A natural way to "generate" these symbols is to linearly process a consecutive chain of received blocks $\mathbf{z}$ with $\mathbf{c}$. The required z_m are then taken from the preceding block. Alternatively, since we transmitted the augmented block $\tilde{\mathbf{x}}$ and have the same channel state at the block boundaries of $\mathbf{z}$, we can describe the transmission process as a mapping from the "ring" of the symbols x_n to the "ring" of the z_n, $n = 0, 1, ..., L-1$. The equalizer maps the z_n to a "ring" of estimates $\hat{x}_n$. We thus joined the block ends of $\mathbf{x}$, $\mathbf{z}$, and $\hat{\mathbf{x}} = [\hat{x}_0\, \hat{x}_1 \cdots \hat{x}_{L-1}]^T$ yielding a "tail-biting" structure. With this interpretation, the equalization step can be denoted as follows:

$$\begin{aligned} \mathbf{p} &\triangleq \mathbf{H}_C^H \mathbf{c}, \\ \hat{\mathbf{x}} &= (\mathbf{Circ_L}[\mathbf{c}])^H \mathbf{z}_0 - (\mathbf{Circ_L}[\mathbf{p}])^H \boldsymbol{\lambda}_0 + \mathbf{p}^H \mathbf{u}\, \boldsymbol{\lambda}_0. \end{aligned} \tag{9}$$

The advantage of this structure becomes obvious if we compute Eq. (9) in the frequency domain by applying the DFT, which is done by multiplying a length L column vector containing elements in the time domain with the $L \times L$ matrix:

$$\mathbf{F} \triangleq (W_{n,k}),\ n, k = 0, 1, ..., L-1,\ W_{n,k} = \exp\left(-\frac{2\pi\, n\, k\, i}{L}\right),$$

where $i \triangleq \sqrt{-1}$. We rewrite Eq. (9) using the property $L \cdot \mathbf{F}^{-1}\mathbf{F} = \mathbf{F}^H\mathbf{F} = L \cdot \mathbf{I}_L$ of the DFT to yield the frequency domain vector $\hat{\mathbf{X}}$ of the estimates $\hat{x}_n$:

$$\begin{aligned} &\mathbf{C} \triangleq \mathbf{F}\,\mathbf{c} = [C_0 \cdots C_{L-1}]^T, \\ &\mathbf{H} \triangleq \mathbf{F}\,[h_0\, h_1 ... h_{M-1}\ \mathbf{0}_{1\times(L-M)}]^T = [H_0 \cdots H_{L-1}]^T, \\ &\mathbf{P} \triangleq \mathbf{F}\,\mathbf{p} = \mathbf{F}\,\mathbf{F}^{-1}\mathbf{Diag}[\mathbf{H}]^H \mathbf{F}\,\mathbf{c} = \mathbf{Diag}[\mathbf{H}]^H \mathbf{C} = [(C_0 \cdot H_0^*) \cdots (C_{L-1} \cdot H_{L-1}^*)]^T, \\ &\hat{\mathbf{X}} = \mathbf{F}\,\hat{\mathbf{x}} = \mathbf{F}(\mathbf{Circ_L}[\mathbf{c}])^H\, \mathbf{F}^{-1}\mathbf{F}\mathbf{z}_0 - \mathbf{F}(\mathbf{Circ_L}[\mathbf{p}])^H\, \mathbf{F}^{-1}\mathbf{F}\boldsymbol{\lambda}_0 + \mathbf{F}\mathbf{p}^H \mathbf{u}\boldsymbol{\lambda}_0 \\ &\quad = \mathbf{Diag}[\mathbf{C}]^H \mathbf{F}\mathbf{z}_0 - \mathbf{Diag}[\mathbf{P}]^H \mathbf{F}\boldsymbol{\lambda}_0 + \frac{1}{L} \cdot \mathbf{F}\mathbf{P}^H \mathbf{F}\,\mathbf{u}\boldsymbol{\lambda}_0 \\ &\quad = \mathbf{Diag}[\mathbf{C}]^H \mathbf{F}\mathbf{z}_0 - \mathbf{Diag}[\mathbf{P}]^H \mathbf{F}\boldsymbol{\lambda}_0 + \frac{1}{L} \cdot \sum_{k=0}^{L-1} (C_k^* \cdot H_k) \cdot \mathbf{F}\boldsymbol{\lambda}_0, \end{aligned} \tag{10}$$

where the 3rd and 5th line follows from the structure of, e.g., $\mathbf{H}_C$ and the last line from $\mathbf{F}\,\mathbf{u} = \mathbf{1}_{L\times 1}$. The vector $\mathbf{C}$ is written in terms of $\mathbf{H}$ using Eq. (7):

$$\begin{aligned} \mathbf{C} = \mathbf{F}\,\mathbf{c} &= \mathbf{F}(\sigma_w^2 \mathbf{I}_L + \mathbf{H}_C\, \mathbf{H}_C^H)^{-1}\, \mathbf{H}_C \mathbf{u} \\ &= \mathbf{F}(\sigma_w^2 \mathbf{I}_L + \mathbf{F}^{-1}\mathbf{Diag}[\mathbf{H}]\mathbf{Diag}[\mathbf{H}]^H \mathbf{F})^{-1}\, \mathbf{F}^{-1}\mathbf{Diag}[\mathbf{H}]\mathbf{F}\mathbf{u} \\ &= (\sigma_w^2 \mathbf{I}_L + \mathbf{Diag}[\mathbf{H}]\mathbf{Diag}[\mathbf{H}]^H)^{-1}\, \mathbf{Diag}[\mathbf{H}]\mathbf{1}_{L\times 1}, \\ \rightarrow C_k &= \frac{H_k}{\sigma_w^2 + H_k\, H_k^*},\ k = 0, 1, ..., L-1. \end{aligned}$$

Given the estimates $\hat{\mathbf{x}} = \mathbf{F}^{-1}\hat{\mathbf{X}}$, the equalizer output $L_e^E(x_n)$ is computed as [12]

$$\mu = \mathbf{c}^H \mathbf{H}_C \mathbf{u} = \mathbf{p}^H \mathbf{u} = \frac{1}{L} \cdot \mathbf{P}^H \mathbf{F}\,\mathbf{u} = \frac{1}{L} \cdot \sum_{k=0}^{L-1} (C_k^* \cdot H_k),$$

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{L} \sum_{n=0}^{L-1} |\mathrm{sign}(\hat{x}_n) \cdot \mu - \hat{x}_n|^2, \\ L_e^E(x_n) &= \frac{2\,\hat{x}_n\, \mu}{\hat{\sigma}^2},\ n = 0, 1, ..., L-1. \end{aligned}$$

```
Input :
    - received symbols $[z_0 \cdots z_{L-1}]^T$, channel and receiver characteristics $h[n]$ and $\sigma_w^2$,
Initialization :
    $[Z_0 \cdots Z_{L-1}]^T \leftarrow \mathrm{DFT}[z_0 \cdots z_{L-1}]^T$,
    $[H_0 \cdots H_{L-1}]^T \leftarrow \mathrm{DFT}[h_0\, h_1 \cdots h_{M-1}\, \mathbf{0}_{1\times(L-M)}]^T$,
    $\mu \leftarrow \frac{1}{L}\sum_{k=0}^{L-1} \frac{H_k H_k^*}{\sigma_w^2 + H_k H_k^*}$,
Equalization :
    $\hat{X}_k \leftarrow \frac{H_k^*}{\sigma_w^2 + H_k H_k^*} Z_k,\ k = 0, 1, ..., L-1$,
    $[\hat{x}_0 \cdots \hat{x}_{L-1}]^T \leftarrow \mathrm{DFT}^{-1}[\hat{X}_0 \cdots \hat{X}_{L-1}]^T$,
    $L_e^E(x_n) \leftarrow \frac{2\,\hat{x}_n}{1-\mu^*},\ n = 0, 1, ..., L-1.$
```

Figure 3: Frequency domain equalizer for "turbo equalization" in absence of a-priori information

```
Input :
    - received symbols $[z_0 \cdots z_{L-1}]^T$, a-priori information $[L_e^D(x_0) \cdots L_e^D(x_{L-1})]^T$,
    - channel and receiver characteristics $h[n]$ and $\sigma_w^2$,
Initialization :
    $[Z_0 \cdots Z_{L-1}]^T \leftarrow \mathrm{DFT}[z_0 \cdots z_{L-1}]^T$,
    $[H_0 \cdots H_{L-1}]^T \leftarrow \mathrm{DFT}[h_0\, h_1 \cdots h_{M-1}\, \mathbf{0}_{1\times(L-M)}]^T$,
    $\mu \leftarrow \frac{1}{L}\sum_{k=0}^{L-1} \frac{H_k H_k^*}{\sigma_w^2 + H_k H_k^*}$,
Equalization :
    $\lambda_n \leftarrow \tanh\left(\frac{1}{2} L_e^D(x_n)\right),\ n = 0, 1, ..., L-1$,
    $[\Lambda_0 \cdots \Lambda_{L-1}]^T \leftarrow \mathrm{DFT}[\lambda_0 \cdots \lambda_{L-1}]^T$,
    $\hat{X}_k \leftarrow \frac{H_k^*}{\sigma_w^2 + H_k H_k^*} Z_k + (\mu - \frac{H_k^* H_k}{\sigma_w^2 + H_k H_k^*}) \Lambda_k,\ k = 0, 1, ..., L-1$,
    $[\hat{x}_0 \cdots \hat{x}_{L-1}]^T \leftarrow \mathrm{DFT}^{-1}[\hat{X}_0 \cdots \hat{X}_{L-1}]^T$,
    $\hat{\sigma}^2 \leftarrow \frac{1}{L}\sum_{n=0}^{L-1} |\mathrm{sign}(\hat{x}_n) \cdot \mu - \hat{x}_n|^2$,
    $L_e^E(x_n) \leftarrow \frac{2\,\hat{x}_n\,\mu}{\hat{\sigma}^2},\ n = 0, 1, ..., L-1.$
```

Figure 4: Frequency domain equalizer for "turbo equalization" using a-priori information

For the first equalization step, no a-priori information is available from the decoder. The equalizer assumes in this case that $L_e^D(x_n) = 0$, $\forall n$, and uses

$$L_e^E(x_n) = \frac{2\,\hat{x}_n}{1 - \mathbf{u}^H \mathbf{H}_C^H \mathbf{c}} = \frac{2\,\hat{x}_n}{1 - \frac{1}{L} \cdot \sum_{k=0}^{L-1} (C_k \cdot H_k^*)},\ n = 0, 1, ..., L-1,$$

to compute $L_e^E(x_n)$ [8, 10]. Figures 3 (first iteration) and 4 (succeeding iterations) depict the complete algorithm to process a block of received symbols z_n to obtain $L_e^E(x_n)$.

3 "Turbo equalization" using soft ISI cancellation

This approach uses soft ISI cancellation [7, 10] instead of equalization to produce estimates $\hat{x}_n$ of the transmitted data x_n. Soft ISI cancellation is performed using a length M FIR filter with coefficients $c_{-k} = h_k^*$, $k = 0, 1, ..., M-1$, matched to the ISI channel response $h[n]$ and a length $2M-1$ FIR feedback filter with coefficients d_k defined as

$$d_0 = 0,\ d_k = \sum_{i=0}^{M-1} h_i\, h_{i-k}^*,\ k = -(M-1), \cdots -1, 1, \cdots, (M-1).$$

Feeding the transmitted symbols x_n into the feedback filter and subtracting the filter output from the matched filter output, the ISI caused of both the channel and the matched filter would be removed completely in the result, the estimate $\hat{x}_n$ [7]. Instead of the inaccessible x_n, the soft bit λ_n, computed according to Eq. (6), is used for feedback. The output LLR $L_e^E(x_n)$ is given as [10]

$$\begin{aligned} \hat{x}_n &= \sum_{k=-M+1}^{0} c_k \, z_{n-k} \; - \sum_{k=-M+1}^{M-1} d_k \, \lambda_{n-k}, \\ L_e^E(x_n) &= \frac{2\,\hat{x}_n}{\sigma_w^2 + \sum_{k=-M+1}^{M-1} (1-\lambda_{n-k}\lambda_{n-k}^*)\, d_k \, d_k^*}, \quad n = 0, 1, ..., L-1. \end{aligned} \tag{11}$$

Applying the same idea as in Section 2, the entire block $\hat{\mathbf{x}}$ of estimates is obtained from the block $\mathbf{z}$ of received symbols and the block $\boldsymbol{\lambda} \triangleq [\lambda_0 \, \lambda_1 \cdots \lambda_{L-1}]^T$ of soft bits as follows:

$$\hat{\mathbf{x}} = \mathbf{H}_C^H \mathbf{z} - \mathbf{H}_C^H \mathbf{H}_C \boldsymbol{\lambda} + \boldsymbol{\lambda}. \tag{12}$$

The first and second term follows from the definition of the c_k (matched filter of $h[n]$) and d_k (cascade of $h[n]$ and its matched filter) and the third term corrects the second term since $d_0 = 0$. Computing Eq. (12) in the frequency domain yields

$$\begin{aligned} \hat{\mathbf{X}} &= \mathbf{F}\,\hat{\mathbf{x}} = \mathbf{F}\,\mathbf{H}_C^H \mathbf{F}^{-1} \mathbf{F}\mathbf{z} - \mathbf{F}\,\mathbf{H}_C\,\mathbf{F}^{-1}\mathbf{F}\,\mathbf{H}_C^H \mathbf{F}^{-1}\mathbf{F}\boldsymbol{\lambda} + \mathbf{F}\boldsymbol{\lambda} \\ &= \mathbf{Diag}[\mathbf{H}]^H \mathbf{F}\mathbf{z} - \mathbf{Diag}[\mathbf{H}]^H \mathbf{Diag}[\mathbf{H}] \mathbf{F}\boldsymbol{\lambda} + \mathbf{F}\boldsymbol{\lambda}. \end{aligned} \tag{13}$$

Applying the inverse DFT to $\hat{\mathbf{X}}$ gives the time domain estimates $\hat{\mathbf{x}}$. For the first equalization with $L_e^D(x_n) = 0$, $\forall n$, we have $\lambda_n = 0$, $\forall n$. The ISI is not reduced at all in this case. It was shown [10] that the SIC is not a suitable equalization algorithm for the first iteration. However, given reasonably good a-priori information (large magnitude $|L_e^D(x_n)|$), the SIC can achieve a BER performance similar to that of MAPEQU [10]. This finding calls for a combined equalization approach, where a more suitable equalization algorithm is used for some initial iterations followed by SIC. To keep the complexity low, we combine appLINEQU and SIC to appLINEQU/SIC.

Figure 5 depicts the complete SIC algorithm using a-priori information $L_e^D(x_n)$ to process a block of received symbols z_n to obtain $L_e^E(x_n)$.

4 Complexity comparison

While the proposed frequency domain algorithms certainly provide savings in computational complexity (for a large ISI channel length M) compared to MAPEQU or LINEQU mentioned in Section 1, we look closer at the complexity of SIC and appLINEQU in their time and frequency domain implementations. To do that, we assume that the soft bits λ_n are already available and skip the computation to obtain the output LLR $L_e^E(x_n)$ (both mappings $\lambda_n \leftarrow L_e^D(x_n)$ and $L_e^E(x_n) \leftarrow \hat{x}_n$ strongly depend on the chosen modulation scheme). Any overhead due to initialization, e.g., to compute $\mathbf{c}$ for time domain appLINEQU or the spectral channel coefficients H_k, is neglected. Also, we assume complex symbols z_n given arbitrary signal constellation alphabets.

Table 2 gives the number of real multiplications and additions per iteration required for equalization of a length L block of symbols z_n. The DFT is carried out using a radix-2 FFT requiring $2L\log_2(L) - 7L + 12$ real multiplications and $3L\log_2(L) - 3L + 4$ real additions for $L = 2^k$, $k = 1, 2, ...$ [13].

```
Input :
    - received symbols $[z_0 \cdots z_{L-1}]^T$, a-priori information $[L_e^D(x_0) \cdots L_e^D(x_{L-1})]^T$,
    - channel and receiver characteristics $h[n]$ and $\sigma_w^2$,
Initialization :
    $d_k \leftarrow \sum_{j=0}^{M-1} h_j\, h_{j-k}^*,\ k = -(M-1), \cdots -1, 1, \cdots, (M-1),\ d_0 = 0,$
    $[\,Z_0 \cdots Z_{L-1}\,]^T \leftarrow \mathrm{DFT}[z_0 \cdots z_{L-1}]^T,$
    $[\,H_0 \cdots H_{L-1}\,]^T \leftarrow \mathrm{DFT}[\,h_0\, h_1 \cdots h_{M-1}\, \mathbf{0}_{1\times(L-M)}\,]^T,$
Soft ISI cancellation :
    $\lambda_n \leftarrow \tanh\left(\frac{1}{2}\, L_e^D(x_n)\right),\ n = 0, 1, ..., L-1,$
    $[\,\Lambda_0 \cdots \Lambda_{L-1}\,]^T \leftarrow \mathrm{DFT}[\,\lambda_0 \cdots \lambda_{L-1}\,]^T,$
    $\hat{X}_k \leftarrow H_k^*\, Z_k + (1 - H_k^* H_k)\, \Lambda_k,\ k = 0, 1, ..., L-1,$
    $[\,\hat{x}_0 \cdots \hat{x}_{L-1}\,]^T \leftarrow \mathrm{DFT}^{-1}[\,\hat{X}_0 \cdots \hat{X}_{L-1}\,]^T,$
    $L_e^E(x_n) \leftarrow \frac{2\,\hat{x}_n}{\sigma_w^2 + \sum_{j=-M+1}^{M-1} (1 - |\lambda_{(n-j) \bmod L}|^2)\, d_j\, d_j^*},\ n = 0, 1, ..., L-1.$
```

Figure 5: Soft ISI cancellation algorithm for "turbo equalization" using a-priori information

approach	real multiplications	real additions
SIC, time domain	$L(8M-8)$	$L(6M-6)$
SIC, frequency domain	$4L\log_2(L) - 6L + 24$	$6L\log_2(L) - 4L + 8$
appLINEQU, time domain	$L(4N+8M)$	$L(4N+2M-2)$
appLINEQU, frequency domain	$4L\log_2(L) - 6L + 24$	$6L\log_2(L) - 4L + 8$

Table 2: Computational complexity of equalization algorithms.

5 Results

We tested the performance of a TEQU-based receiver using MAPEQU and SIC or appLINEQU (both in their frequency domain implementation) through simulation. Performance criterion is the BER if data is encoded to length L blocks of code symbols c_n and transmitted (after interleaving and augmentation) over a length 5 ISI channel with response $h[n] = 0.227\,\delta[n] + 0.46\,\delta[n-1] + 0.688\,\delta[n-2] + 0.46\,\delta[n-3] + 0.227\,\delta[n-4]$. The used recursive systematic convolutional encoder has the generator $G(D) = (1, \frac{1+D^2}{1+D+D^2})$. The noise variance σ_w^2 is determined according to the SNR $\frac{E\{|z_n|^2\}}{N_0} = \frac{E\{|x_n|^2\}}{N_0} = \frac{1}{2\,\sigma_w^2}$, where we used the fact that the response $h[n]$ has unit energy, i.e., $\sum_{k=0}^{4} |h_k|^2 = 1$. The time and frequency domain implementations of SIC yield identical performance results. For appLINEQU, the frequency domain implementation equals a time domain implementation with the filter parameters $N_1 = L-1$ and $N_2 = 0$. Approximate similarity is still given for a sufficiently large filter length N. In [10], $N_1 = 9$ and $N_2 = 5$ ($N = 15$) was used for the same $h[n]$, which we think is a reasonable choice for this channel.

Figure 6 shows the BER performance after transmission of 5000 blocks ($L = 2048$), where the receiver performed 20 iterations. References are the performance of coded data transmission over an AWGN channel at the same SNR with MAP-based decoding and one-time MAPEQU and MAP-based decoding. The combined appLINEQU/SIC approach uses appLINEQU for the first 15 iterations. Insight in how to choose which equalization approach for how many iterations is given in [10]. Note that the rather large number of iterations is necessary only for SNRs close to the minimum SNR with significant BER improvement over the iterations ("water fall" region). We refer to Bauch [14], who applied several iteration termination criteria for iterative algorithms to TEQU, for further information.

Figure 7 shows the computational complexity of SIC and appLINEQU in their time and frequency implementations obtained using Table 2 and the system parameters M and N as function of L.

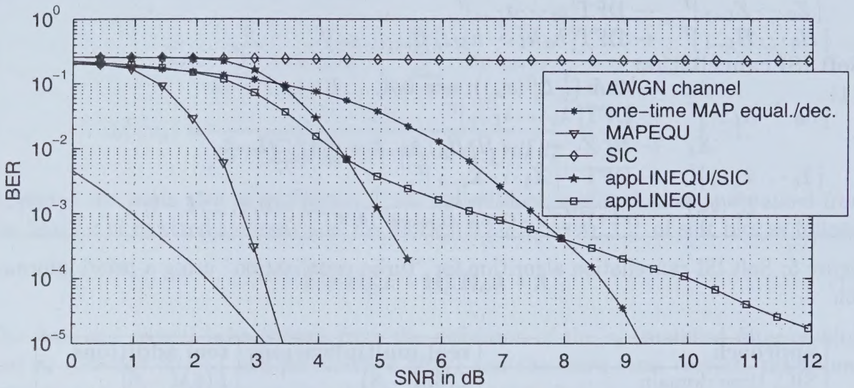


Figure 6: BER performance of "turbo-equalization"-based receivers.

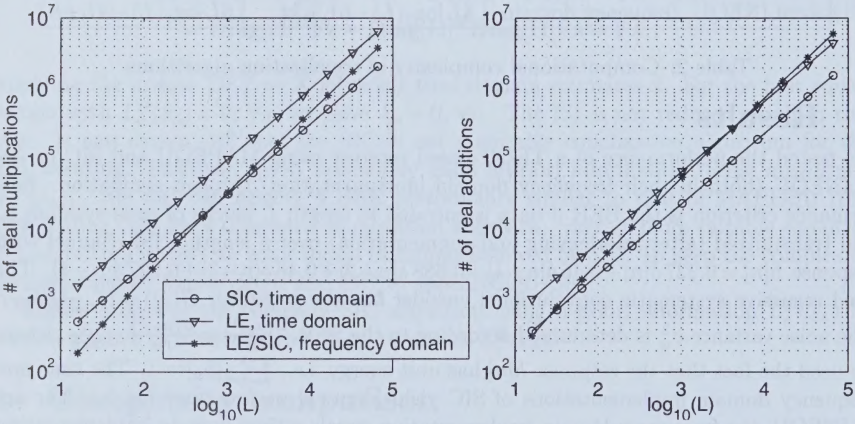


Figure 7: Computational complexity of SIC and appLINEQU (LE).

6 Conclusions

The results confirm the remarkable performance of "turbo equalization" in general and especially of the introduced equalization algorithms, which were rederived from its time domain original to perform equalization in the frequency domain. With the iterative receiver algorithm, the degradation due to ISI (compare the performance using the BER-optimal MAP-based algorithms one time) can be reduced significantly. The surprisingly good performance of appLINEQU/SIC developed in [10] despite the poor performance of pure SIC or appLINEQU, if applied to TEQU, calls for more research on these hybrid equalization schemes. The complexity comparison in Figure 7 proves that the frequency domain versions of SIC and appLINEQU provide savings in complexity up to a threshold

L, what suggests the use of these algorithms for short block lengths. This threshold increases with larger ISI channel response length M (SIC, appLINEQU) and filter length N (appLINEQU).

References

[1] J. Hagenauer, "The turbo principle: Tutorial introduction and state of the art," *Intern. Symposium on Turbo Codes*, pp. 1–11, Sep 1997.

[2] C. Berrou and A. Glavieux, "Near optimum error correcting coding and decoding: Turbo codes," *IEEE Trans. on Comm.*, vol. 44, pp. 1261–1271, October 1996.

[3] C. Berrou, A. Glavieux, and P. Thitimajshima, "Near shannon limit error-correcting coding and decoding: Turbo codes," *Proc. of ICC '93, Geneva*, May 1993.

[4] C. Douillard et al., "Iterative correction of intersymbol interference: Turbo equalization," *European Trans. on Telecomm.*, vol. 6, pp. 507–511, Sep-Oct 1995.

[5] G. Bauch and V. Franz, "A comparison of soft-in/soft-out algorithms for "turbo detection"," *Proc. on the Intern. Conference on Telecomm. (ICT '98)*, pp. 259–263, June 1998.

[6] A. Anastasopoulos and K. Chugg, "Iterative equalization/decoding for TCM for frequency-selective fading channels," *Conference Record on the 31th Asilomar Conference on Signals, Systems & Computers*, vol. 1, pp. 177–181, November 1997.

[7] A. Glavieux, C. Laot, and J. Labat, "Turbo equalization over a frequency selective channel," *International Symposium on Turbo codes & related topics*, pp. 96–102, September 1997.

[8] M. Tüchler, "Iterative equalization using priors," Master's thesis, University of Illinois at Urbana-Champaign, U.S.A., 2000.

[9] R. Graham, D. Knuth, and O. Patashnik, *Concrete Mathematics, 2nd Edition.* Reading, Massachusetts: Addison-Wesley, 1994.

[10] M. Tüchler, R. Kötter, and A. Singer, "Turbo equalization: principles and new results," *submitted to IEEE Trans. on Communications*, 2000.

[11] L.R. Bahl et al., "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. on Information Theory*, vol. 20, pp. 284–287, March 1974.

[12] M. Tüchler, A. Singer, and R. Kötter, "Minimum mean squared error (MMSE) equalization using priors," *submitted to IEEE Trans. on Signal Processing*, 2000.

[13] D. Jones, *Class notes for ECE451: Advanced Digital Signal Processing.* Urbana, IL, U.S.A.: University of Illinois at Urbana/Champaign, Fall 1998.

[14] G. Bauch, H. Khorram, J. Hagenauer, "Iterative equalization and decoding in mobile communications systems," *Second European Personal Mobile Communications Conference (2. EPMCC '97)*, pp. 307–312, Sep/Oct 1997.

Multimodal Human-Computer Interfaces[*]

J. L. Flanagan
CAIP Center
Rutgers University
96 Frelinghuysen Road, CoRE #706, Piscataway, NJ 08854-8088
Email: jlf@caip.rutgers.edu

Abstract

User interfaces that capture the ease of natural communication may be key to mass deployment of networked information systems. Humans favor the sensory modes of sound, sight and touch for information exchange. And, interfaces that accommodate these modes can transcend both the utility and functionality of traditional mouse and keyboard. Technologies that support these aspirations are as yet primitive. But, applied prudently, they can be made to perform advantageously. This report gives a perspective on one research effort aimed to fuse simultaneous sound, sight and touch interaction, utilizing automatic speech recognition, text-to-speech synthesis, hands-free sound capture, visual gesture and eye tracking, manual gesture and force feedback grasp, and multimodal semantic analysis.

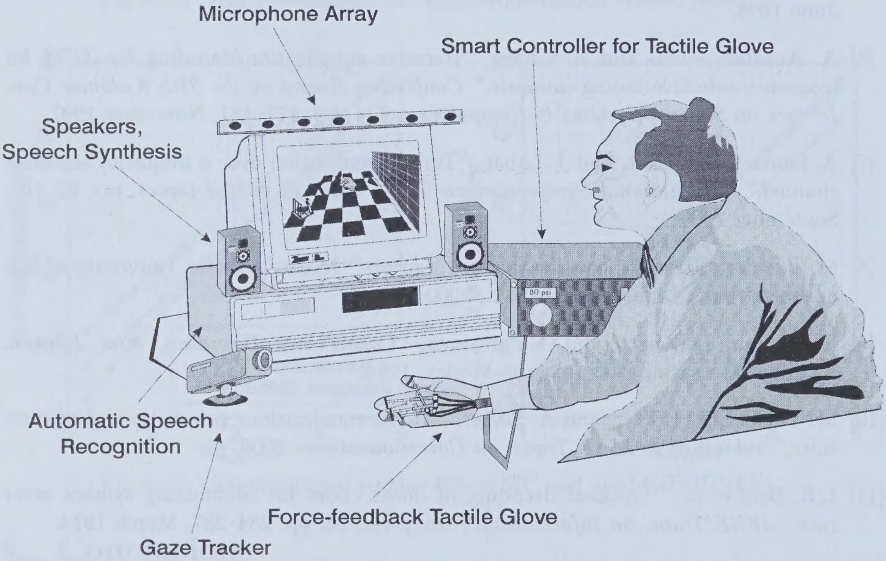

Microphone Array
Smart Controller for Tactile Glove
Speakers,
Speech Synthesis
80 psi
Automatic Speech
Recognition
Force-feedback Tactile Glove
Gaze Tracker

Fig. 1. Concept of computer user station incorporating sight, sound and touch modalities for interaction. An eye tracker provides a screen cursor at the location to which gaze is directed. A gimbaled camera, an ultrasonic ranger, and a co-located infrared source locate the face and

eye, illuminate the eye, and compute the angle between the corneal reflection and the centroid of the pupil. A microphone array on the workstation is fix-focused on the user position and captures good quality sound in noisy environments. User speech commands are delivered to an automatic speech recognizer based upon principles of hidden Markov models. Machine answerback is effected by text-to-speech synthesis computed by the system. A force feedback tactile glove, whose spatial position is determined from a Polhemus coil, enables control of displayed objects by manual gesture and by grasp. Intelligent software agents fuse the simultaneous sensory inputs, making rudimentary semantic analyses that support task-awareness and estimates of user intent [1,2].

---

[*] Components of this research are supported by NSF Contracts #IRI-96-18854 (STIMULATE) and #IIS-98-72995 (KDI), DARPA Contract #N66001-96-C-8510 (DISCIPLE) and by Intel Corp. and Texas Instruments.

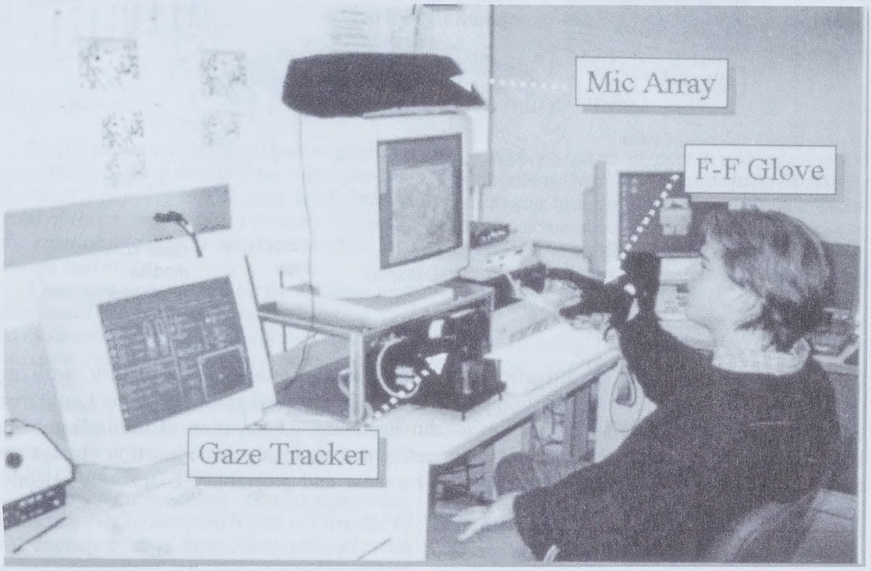


Fig. 2. Multimodal user station implemented in the laboratory. The interface incorporates simultaneous modalities for sound, sight, and touch, and intelligent software agents for interpreting and fusing these signals [1, 2].

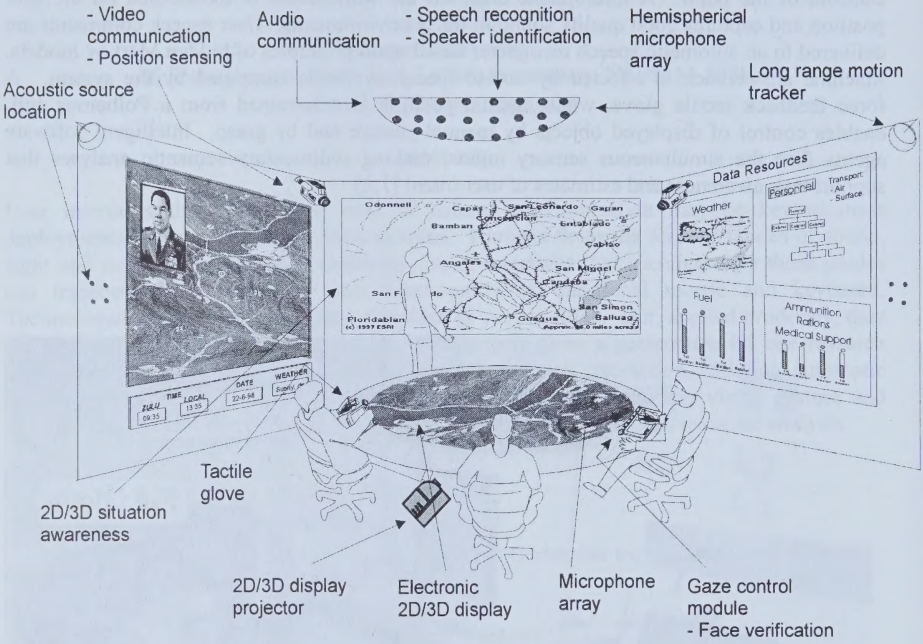

- Video communication
- Position sensing
Audio communication
- Speech recognition
- Speaker identification
Hemispherical microphone array
Long range motion tracker
Acoustic source location
Data Resources
Personnel
Transport: - Air - Surface
Weather
Fuel
Ammunition
Rations
Medical Support
ZULU
TIME
LOCAL
DATE
WEATHER
Tactile glove
2D/3D situation awareness
2D/3D display projector
Electronic 2D/3D display
Microphone array
Gaze control module
- Face verification

Fig. 3. Conceptualized multiparticipant audio/video conferencing facility utilizing multimodal interfaces, 2Dand 3D graphics, and networked databases. Individual conferees are equipped for multimodal interaction as shown in Fig. 2. Position of multiple moving talkers is determined by combined video processing and acoustic source location. Image and sound capture are effected by slaved video cameras and steerable 3D-spatially-selective microphone array [3].

References:

1. J. Flanagan, G. Burdea, C. Kulikowski, J. Wilder, I. Marsic, Synergistic Multimodal Communication in Collaborative Multiuser Environments, NSF Final Report, Contract #9618854 (STIMULATE), March 31, 2000. see also http://www.nsf.gov/od/lpa/news/press/00/pr0051.htm
2. I. Marsic, A. Medl, J. Flanagan, Natural Communication with Information Systems, Proc. IEEE, special issue on speech (in press).
3. J. Flanagan, Autodirective Sound Capture: Towards Smarter Conference Rooms, IEEE Intelligent Systems, vol. 14, pp. 16-19, Mar./Apr. 1999.

Synthetic Talking Face Driven by Multimodal Signal

Shigeo MORISHIMA
Seikei University
3-3-1 Kichijiji-kitamachi Musashino Tokyo 180-8633 Japan
shigeo@ee.seikei.ac.jp

ATR Media Integration and Communications Research Laboratories
2-2-2 Hikaridai Seikacho Soraku-gun Kyoto 619-0288 Japan
shigeo@mic.atr.co.jp

Abstract

In this paper, we describe how to generate a synthetic talking face in a real-time process by exactly copying a real person's face. The goal of this research is to precisely duplicate the emotion and impression of an original face image and voice.

Firstly, a semiautomatic face fitting tool from multi-angle camera images is introduced to customize a realistic 3-D face model with texture and geometry very close to the original.

Secondly, a real-time mouth shape control mechanism is proposed by using multi-layer neural network to convert speech parameters to lip shape parameters.

For dynamic modeling of facial expression, a muscle structure constraint is introduced to generate a facial expression naturally with a few parameters. And this is used to obtain muscle parameters automatically for setting the expression from a local motion vector on the face calculated by the optical flow of a video sequence. Furthermore, we also tried to control this artificial muscle model directly by EMG signal.

To get greater reality, a modeling method of hair is also introduced, and the dynamics of hair in a wind stream can be achieved at low calculation cost. By using these several kinds of multi-modal signal sources, a very natural face image and impression can be duplicated on an avatar's face.

1. Introduction

A very important factor in making an avatar look believable or alive depends on how precisely an avatar can duplicate a real human facial expression and impression on a face precisely. Especially in the case of communication applications using avatars, real-time processing with low delay is essential.

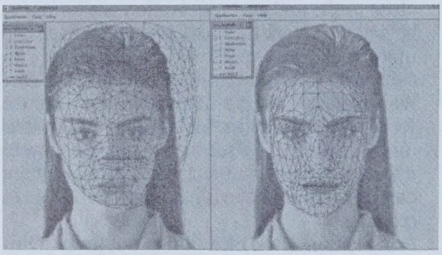
Our final goal is to generate a virtual space close to the real communication environment between network users or between humans and machines. There should be an avatar in cyberspace that projects the features of each user with a realistic texture-mapped face to generate facial expression and action controlled by a multimodal input signal. Users can also get a view in cyberspace through the avatar's eyes, so they can communicate with each other by gaze crossing.

In Section 2, the face fitting tool from multi-angle camera images is introduced to make a realistic 3-D face model with texture and geometry very close to the original. This fitting tool is a GUI-based system using easy mouse operation to pick up each feature point on a face contour, and the face parts can enable easy construction of a 3-D personal face model.

When an avatar is speaking, the voice signal is essential in determining the mouth shape feature. Therefore, a real-time mouth shape control mechanism is proposed by using a neural network to convert speech parameters to lip shape parameters. This neural network can realize an interpolation between specific mouth shapes given as learning data (Morishima and Harashima 1991, Morishima 1997). The emotional factor can sometimes be captured by speech parameters. This media conversion mechanism is described in Section 3.

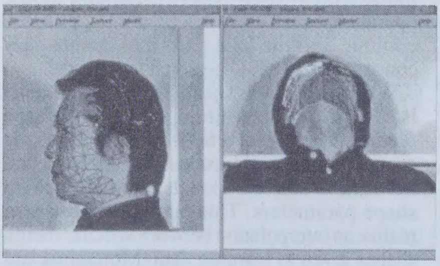
For dynamic modeling of facial expresson, a muscle structure constraint is introduced for making a facial expression naturally with few parameters.We also tried to obtain muscle parameters automatically from a local motion vector on the face calculated by the optical flow in a video sequence. Accordingly, 3-D facial expression transition is duplicated from actual people by analysis of video images captured by a 2-D camera without landmarks on the face. These efforts are described in Section 4. Furthermore, we tried to control this artificial muscle direcly by EMG signal processing.

To get greater reality for the head, a modeling method of hair is introduced and the dynamics of hair in a wind stream can be achieved at low calculation cost. This is explained in Section 5. By using these several kinds of multi-modal signal sources, a very natural face image and impression can be duplicated on an avatar's face.



a) Initial model b) Fitted model

Fig.1 Frontal model fitting by GUI tool



a) Profile view b) Bottom view

Fig.2 Multi-view fitting tool

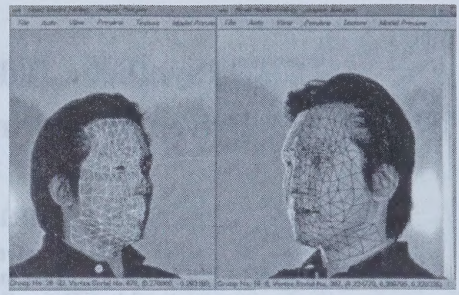


Fig.3 Multi-view fitting to oblique angle

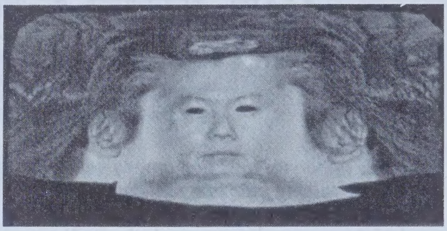
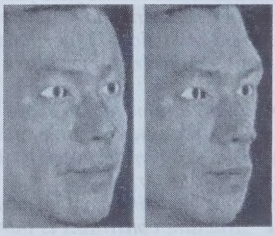


Fig.4 Cylindrical face texture



a) Multi-view b) Two-view

Fig.5 Reconstructed face

2. Face Modeling

To generate a realistic avatar's face, a generic face model is manually adjusted to the user's face image. To produce a personal 3-D face model, both the user's frontal face image and profile image are necessary at the least. The generic face model has all of the control rules for facial expressions defined by the FACS parameter as a 3-D movement of grid points to modify geometry.

Figure 1 shows a personal model both before and after the fitting process for a front-view image by using our original GUI based face fitting tool. The front view and profile images are put into the system and then corresponding control points are manually moved to a reasonable position by mouse operation.

The synthesized face emerges by mapping of the blended texture generated by the user's frontal image and profile image onto the modified personal face model. However, sometimes self-occlusion occurs and then we cannot capture texture from only the front and profile face images in the occluded part of the face model. To construct a 3-D model more accurately, we also introduce a multi-view face image fitting tool.

Figure 2(b) shows the fitting result to the captured image from the bottom angle to compensate for the texture behind the jaw. Figure 3 shows the fitting result with the face image from any oblique angle. The lotation angle of the face model can be controlled in the GUI preview window to achieve the best fitting to the face image captured from any arbitrary angle. Figure 4 shows the full face texture projected onto a cylindrical coordinate. This texture is projected onto a 3-D personal model adjusted to fit to multi-view images to synthesize a face image. Figure 5 shows examples of reconstructed faces. Figure 5(a) uses nine view images and Figure 5(b) uses only frontal and profile views. Much better image quality can be achieved by the multi-view fitting process.

3. Voice Driven Talking Avatar

To realize lip synchronization, the spoken voice is analyzed and converted into mouth shape parameters by a neural network on a frame-by-frame basis.

Multiple-user communication systems in cyberspace are constructed based on a server-client system. In our system, only a few parameters and the voice signal are transmitted through the network. The avatar's face is synthesized by these parameters locally in the client system.

3.1 Parameter Conversion

In the server system, the voice from each client is phonetically analyzed and converted to mouth shape and expression parameters.

LPC Cepstrum parameters are converted into mouth shape parameters by a neural network trained by vowel features. Figure 6 shows the neural network structure used for parameter conversion. Twenty dimensional Cepstrum parameters are calculated every 32 ms with 32 ms frame length. In the client system, the on-line captured voice of each user is digitized a 16 KHz and 16 bits and transmitted to the server system frame-by-frame through a network in the communication system. Then the mouth shape of each avatar in cyberspace is synthesized by this mouth shape parameter received at each client.

The emotion condition is also determined by

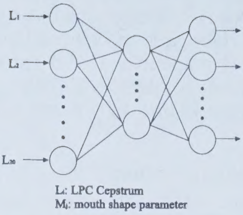


Fig.6 Network for parameter conversion

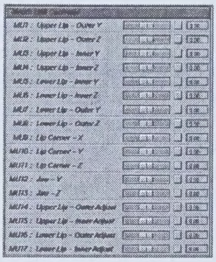
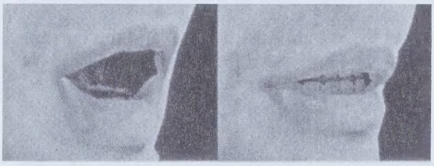
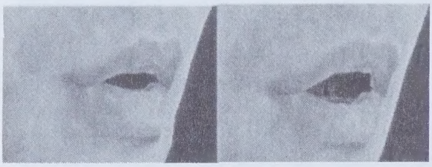

MU1 : Upper Lip - Outer Y
MU2 : Upper Lip - Outer Z
MU3 : Upper Lip - Inner Y
MU4 : Upper Lip - Inner Z
MU5 : Lower Lip - Inner Y
MU6 : Lower Lip - Inner Z
MU7 : Lower Lip - Outer Y
MU8 : Lower Lip - Outer Z
MU9 : Lip Corner - X
MU10 : Lip Corner - Y
MU11 : Lip Corner - Z
MU12 : Jaw - Y
MU13 : Jaw - Z
MU14 : Upper Lip - Outer Adjust
MU15 : Upper Lip - Inner Adjust
MU16 : Lower Lip - Outer Adjust
MU17 : Lower Lip - Inner Adjust

Fig.7 Control panel for mouth shape editor



a) Shape for /a/ **b) Shape for /i/**



c) Shape for /u/ **d) Shape for /o/**

Fig.8 Typical mouth shapes

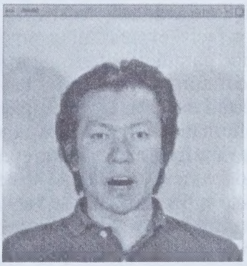


Fig.9 Synthesized face speaking /a/

LPC Cepstrum, power, pitch frequency, and utterance speed by using discriminant analysis into Anger, Happiness or Sadness. Each basic emotion has a specific facial expression parameter described by FACS (Facial Action Coding System, Ekman and Friesen 1978).

3.2 Mouth Shape Editor

Mouth shape can be easily edited by our mouth shape editor (Figure 7). We can change each mouth parameter to set a specific mouth shape on the preview window. Typical vowel mouth shapes are shown in Figure 8. Our special mouth model has polygons for inside the mouth and the teeth. A tongue model is now under construction. For parameter conversion from LPC Cepstrum to mouth shape, only the mouth shapes for five vowels and nasals are defined as the training set. We have defined all of the mouth shapes for Japanese phonemes and English phonemes by using this mouth shape editor. Figure 9 shows a synthesized avatar face speaking phoneme /a/.

3.3 Multiple User Communication System

Each user can walk through and fly through cyberspace by mouse control and the current locations of all users are always monitored by the server system. The avatar image is generated in a local client machine by the location information from the server system.

The emotion condition can always be determined by voice, but sometimes an user gives his or her avatar a specific emotion condition by pushing a function key. This process works as first priority. For example, push anger and then the angry face of your avatar emerges.

The location information of each avatar, mouth shape parameters, and emotion parameters are transmitted every 1/30 seconds to the client systems. The distance between any two users is calculated by the avatar location information, and the voice from every user except himself or herself is mixed and amplified with gain according to the distance. Therefore, the voice from the nearest avatar is very loud and that from far away is silent.

Based on facial expression parameters and mouth shape parameters, an avatar's face is synthesized frame-by-frame. Also, the avatar's body is located in cyberspace according to the location information. There are two modes for display: the view from avatar's own eyes for eye contact and the view from the sky to search for other users in cyberspace. These views can be chosen by a menu in the window. Figure 10 shows a communication system in cyberspace with an avatar.

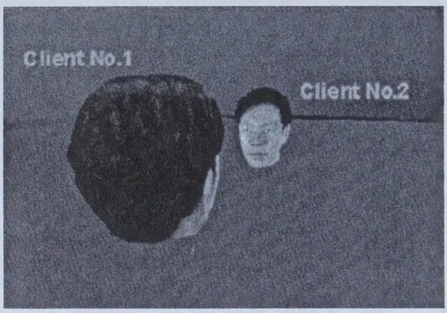


Fig.10 Communication system in cyberspace

3.4 User Adaptation

When a new user enters the system, his or her face model and voice model have to be registered before operation. In case of voice, ideally new learning for the neural network should be performed. However, it takes a very long time to get convergence of back-propagation. To simplify the face model construction and voice learning, a GUI tool for speaker adaptation has been prepared. To register the face of a new user, a generic 3D face model is modified to fit on the input face image. Expression control rules are defined in the generic model, so every user's face can be equally modified to generate basic expressions using the FACS-based expression control mechanism.

For voice adaptation, 75 persons' voice data including five vowels are pre-captured, and a database for weights of the neural network and voice parameters is constructed. Accordingly, speaker adaptation is performed by choosing the optimum weight from the database. Training of the neural network for the data of each of the every 75 persons is already finished before operation. When a new non-registered speaker comes in, he or she has to speak five vowels into a microphone. LPC Cepstrum is calculated for every each of the five vowels, and this is entered into the neural network. Then the mouth shape is calculated by the selected weight and the error between the true mouth shape and the generated mouth shape is evaluated. This process is applied to the entire database one-by-one and the optimum weight is selected when the minimum error is detected.

3.5 System Features

Natural communication environment between multiple users in cyberspace by transmission of natural voice and real-time synthesis of an avatar's facial expression is realized. The current system works with three users in an intra-network environment at the speed of 16 frames per second on an SGI Max IMPACT. An emotion model (Morishima 1996) is introduced to improve the communication environment. Gaze tracking and mouth closing point detection can also be realized by a pan-tilt-zoom controlled camera.

3.6 Entertainment Application

When people watch movies, they sometimes overlap their own figure with the actor's image. An interactive movie system we constructed is an image creating system in which the user can control facial expression and lip motion of his or her face image inserted into a movie scene. The user submits a voice sample by microphone and pushes keys that determine expression and special effect. His or her own video program can be generated in real-time.

At first, once a frontal face image of a visitor is captured by camera, a 3-D generic wireframe model is fitted onto the user's face image to generate a personal 3-D surface model. A facial expression is synthesized by controlling the grid point of the face model and texture mapping. For speaker adaptation, the visitor has to speak five vowels to choose an optimum weight from the database.

In the interactive process, a famous movie scene is going on and the facial region of an actor or actress is replaced with visitor's face. Facial expression and lip shape are also controlled synchronously by the captured voice. An active camera tracks the visitor's face, and the facial expression is controlled by a CV -based face image analysis.Figure 11 shows the original movie clip, and Figure 12 shows the result of fitting a face model into this scene. Figure 13 shows an user's face inserted into actor's face after color correction. Any expression can be appended and any scenario can be given by the user independent of the original story in this interactive movie system.

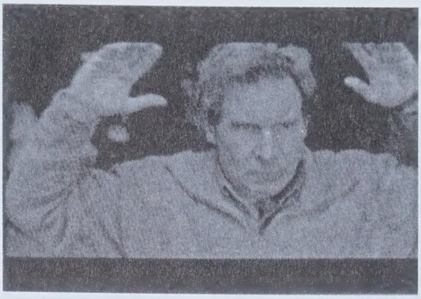


Fig.11 Original movie clip

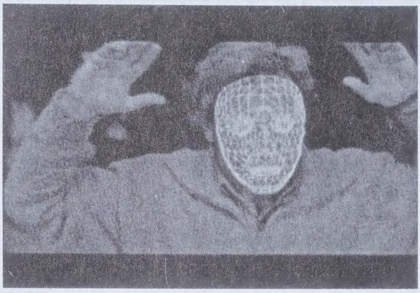


Fig.12 Fitting result of face model

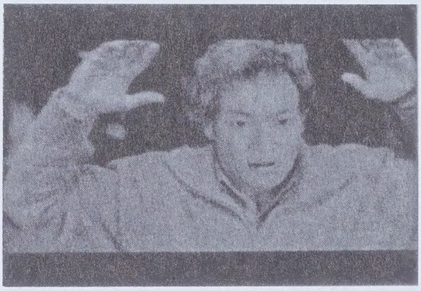


Fig.13 Reconstructed face by user's

4. Muscle Constraint Face Model

Muscle based face image synthesis is one of the most realistic approaches to the realization of a life-like agent in computers. A facial muscle model is composed of facial tissue elements and simulated muscles. In this model, forces are calculated to effect a facial tissue element by contraction of each muscle string, so the

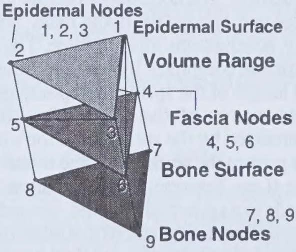


Fig.14 Layered tissue element

combination of each muscle's contracting force decides a specific facial expression. This muscle parameter is determined on a trial-and-error basis by comparing a sample photograph and a generated image using our Muscle-Editor to generate a specific face image. In this section, we propose the strategy of automatic estimation of facial muscle parameters from 2-D optical flow by using a neural network. This corresponds to the 3-D facial motion capturing from 2-D camera images under a physics-based condition without markers.

We introduce a multi-layer back-propagation network for the estimation of the muscle parameter. The facial expression is then re-synthesized from the estimated muscle parameter to evaluate how well the impression of an original expression can be recovered. We also tried to generate animation by using the captured data from an image sequence. As a result, we can obtain and synthesize an image sequence that gives an impression very close to the original video.

4.1 Layered Dynamic Tissue Model

The human skull is covered by a deformable tissue that has five distinct layers. Four layers (epidermis, dermis, subcutaneous connective tissue, and fascia) comprise the skin, and the fifth layer comprises the muscles responsible for facial expression. In accordance with the structure of real skin, we employ a synthetic tissue model constructed from the elements illustrated in Figure 14, consisting of nodes interconnected by deformable springs (the lines in the figure). The epidermal surface is defined by nodes 1, 2, and 3, which are connected by epidermal springs. The epidermal nodes are also connected by dermal-fatty layer springs to nodes 4, 5, and 6, which define the fascia surface. Fascia nodes are interconnected by fascia springs. They are also connected by muscle layer springs to skull surface nodes 7, 8, 9. The facial tissue model is implemented as a collection of node and spring data structures. The node data structure includes variables to represent the nodal mass, position, velocity, acceleration, and net force. The spring data structure comprises the spring stiffness, the natural length of the spring, and pointers to the data structures of the two nodes that are interconnected by the spring. Newton's laws of motion govern the response of the tissue model to force (Lee, Terzopoulos and Waters 1995). This leads to a system of coupled, second-order ordinary differential equations that relate the node positions, velocities, and accelerations to the nodal forces.

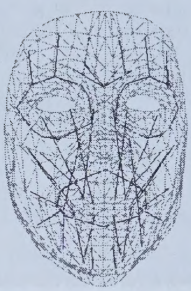
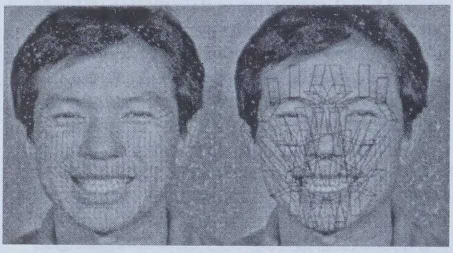


Fig.15 Facial muscle model



a) Optical flow **b) Feature window**

Fig.16 Motion feature vector on face

4.2 Facial Muscle Model

Figure 15 shows our simulated muslces. The black line indicates the location of each facial muscle in a layered tissue face model. Each muscle is modeled by a combination of separated muscles, and the left and right parts of the face are symmetric. For example, the frontalis has a triangularly spread feature, so three kinds of linear muscles model this frontalis on one side. The frontalis pulls up the eyebrows and makes wrinkles in the forehead. The corrugater, which is modeled by four springs located between the eyes, pulls the eyebrows together and makes wrinkles between the left and right eyebrows. However, those muscles can't pull down the eyebrows and make the eyes thin. Therefore, the orbicularis oculi muscle is also appended in our model. The orbicularis oculi is separated into an upper part and a lower part. To make muscle control simple around the eye area, the orbicularis oculi is modeled with a single function in our model. Normally, muscles are located between a bone node and a fascia node. But the Orbicularis oculi has an irregular style, whereby it is attached between fascia nodes in a ring configuration; it has eight linear muscles that approximate a ring muscle. Contraction of the

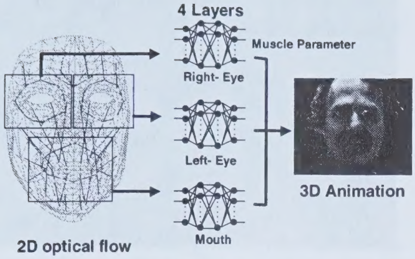
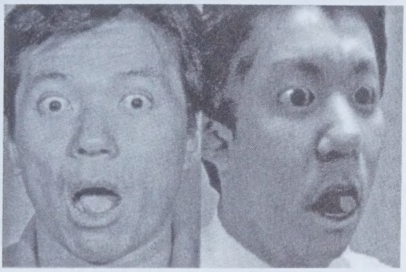


Fig.17 Conversion from motion to animation



a) Original b) Regenerated

Fig.18 Face expression regeneration

ring muscle makes the eyes thin. The muscles around the mouth are very important for the production of speaking scenes. Most Japanese speaking scenes are composed of vowels, so we mainly focused on the reproduction of vowel mouth shapes as a first step and relocated the muscles around the mouth (Sera, Morishima and Terzopoulos 1996). As a result, the final facial muscle model has 14 muscle springs in the forehead area and 27 muscle springs in the mouth area.

4.3 Motion Capture on Face

A personal face model is constructed by fitting the generic control model to personal range data. An optical flow vector is calculated in a video sequence, and we accumulate the motion in the sequence from neutral to each expression. Then motion vectors are averaged in each window as shown in Figure 16(b). This window location is determined automatically in each video frame.

A layered neural network finds a mapping from the motion vector to the muscle parameter. A four-layer structure is chosen to effectively model the nonlinear performance. The first layer is the input layer, which corresponds to a 2-D motion vector. The second and third layers are hidden layers. Units of the second layer have a linear function, and those of the third layer have a sigmoid function. The fourth layer is the output layer, corresponding to the muscle parameter, and has linear units. Linear functions in the input and output layers are introduced to maintain the range of input and output values.

A simpler neural network structure can help to speed up the convergence process in learning and reduce the calculation cost in parameter mapping, so the face area is divided into three sub-areas as shown in Figure 17. They are the mouth area, left-eye area, and right-eye area, each giving independent skin motion. Three independent neural networks are prepared for these three areas.

Learning patterns are composed of basic expressions and the combination of their muscle contraction forces. We created six basic facial expressions consisting of anger, disgust, fear, happiness, sadness and surprise. Basic mouth shapes were made for vowels "a", "i", "u", "e" and "o", and a closed mouth shape.

A motion vector is given to a neural network from each video frame, and then a facial animation is generated from the output muscle parameter sequence. This is a test of the effect of interpolation on our parameter conversion method based on the generalization of the neural network.

Figure 18 shows the result of expression regeneration for surprise from an original video sequence.

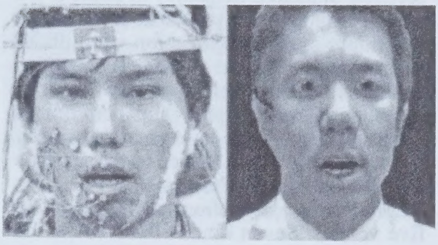


Fig.19 Mouth shape control by EMG

4.4 Muscle Control by EMG

EMG data is the voltage wave captured by a needle inserted directly into each muscle so that the feature of wave expresses the contraction situation of a muscle. In particular, seven di-ball wires were inserted into the muscles around the mouth to make a model of mouth shape control.

The conversion from EMG data to muscle parameters is as follows. First, the average power of an EMG wave is calculated every 1/30 seconds. Sampling frequency of EMG is 2.5 kHz. The maximum power is equal to the maximum muscle contraction strength by normalization. Then the converted contraction strength of each muscle is given every 1/30 seconds and then the facial animation scene is generated. However, the jaw is controlled directly by the marker position located on the subject's jaw independent of EMG data.

EMG data is captured from a subject speaking five vowels. Seven kinds of waves can be captured and seven contractions of muscles are determined. Then the face model is modified and the mouth shape for each vowel is synthesized. The result is shown in Figure 19. In Figure 19, the camera captured image and synthesized image can be compared, and the impression is very close each other. Next, EMG time sequence is converted into muscle parameters every 1/30 seconds and facial animation is generated. Each sample is composed of a three-second sentence. Good animation quality for speaking the sentence is achieved by picking up impulses from the EMG signal and activating each appropriate muscle. This data offer the precise transition feature of each muscle contraction.

5. Dynamic Hair Modeling

The naturalness of hair is a very important factor in a visual impression, but it is treated with a very simple model or as a texture. Because real hair has huge pieces and complex features, it's one of the hardest targets to model by computer graphics. In this section, a new hair modelling system is presented. This system helps the designer to create any arbitrary hair style by computer graphics with easy operation by using an editor of tufts. Each piece of hair has an independent model, and dynamic motion can be easily simulated by solving a motion equation. However, each piece has only a few segments and the feature is modelled by a 3-D B-spline curve, so the calculation cost is not so huge.

5.1 Modelling Hair

It is necessary to create a model of each piece of hair to generate a natural dynamic motion when wind is flowing. However, the hair feature is very complex and a real head has more than 100,000 pieces. Therefore, a polygon model for hair needs a huge amount of memory for the entire head to be stored. In this paper, each piece has only a few segments to control and is expressed with a 3D B-Spline curve.

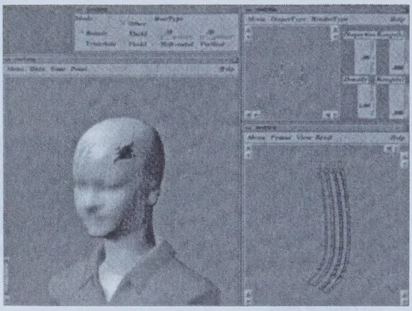


Fig.20 GUI tool for hair designing

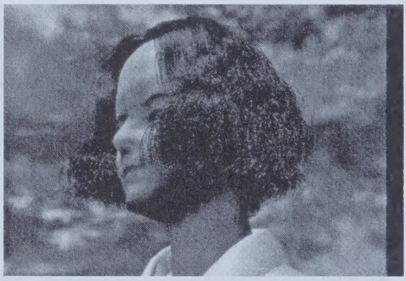


Fig.21 Example of hair style

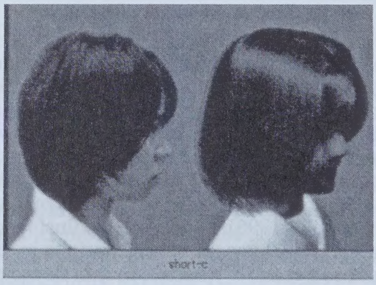


Fig.22 Copying real hair style

5.2 Tuft Model

In our designing system, a head is composed of about 3400 polygons and a 3-D B-Spline curve comes out from each polygon to create hair style. The hair generation area is composed of about 1300 polygons, and one specific hair feature is generated and copied for each polygon. To manipulate hair to get a specific hair style, some of the hair features are simultaneously treated as one tuft. Each tuft is composed of more than seven squares in which hair segments pass through. This square is gathering control point of the hair segments, so manipulation of this square can realize any hair style by rotation, shift and modification.

5.3 Hair Style Designing System

There are three processes to generate hair style by GUI, and each process is easily operated by mouse control. The first process is to decide the initial region on the head from which the tuft comes out. The second process is to manipulate a hair tuft by modifying the squares by rotation, shift and expansion. The third process is matching a tuft onto the surface of the head. After these processes, each polygon gets one hair feature, and in total 1300 features are available. Finally, 20,000 to 30,000 pieces of hair for static images, or 100,000 to 200,000 pieces for animation, are generated by copying the feature in each polygon. The GUI tool for hair designing is shown in Figure 20.

5.4 Rendering

Each hair piece is modelled with a 3-D B-Spline curve and also modelled as a very thin circular pipe. Therefore, the surface normal can be determined in each pixel to introduce the Lambert model and Phong model. The typical hair style "Dole Bob" is shown in Fig. 21. Figure 22 shows an example modeling of a real hair style. The impression of the generated hair is very natural.

6. Conclusions

To generate a realistic avatar's face for face-to-face communication, a multi-modal signal is introduced to duplicate original facial expressions. Voice is used to realize lip synchronization and expression control. Video captured images are used to regenerate an original facial expression under a facial muscle constraint. EMG data is used to directly control an artificial muscle. Finally, a hair modeling method is proposed to make an avatar appear more natural and believable.

References

Morishima, S. and Harashima, H. (1991). A Media Conversion from Speech to Facial Image for Intelligent Man-Machine Interface, IEEE JSAC, Vol. 9, No. 4, pp. 594-600.

Morishima, S. (1997). Virtual Face-to-Face Communication Driven by Voice Through Network. Workshop on Perceptual User Interfaces, pp. 85-86.

Ekman, P. and Friesen, W.V. (1978). Facial Action Coding System. Consulting Psychologists Press Inc.

Essa, I., Darrell, T. and Pentland, A. (1994). Tracking Facial Motion. Proceedings of Workshop on Motion and Nonrigid and Articulated Objects, pp.36-42.

Mase, K. (1991). Recognition of Facial Expression from Optical Flow. IEICE Transactions, Vol E 74, No. 10.

Morishima, S. (1996). Modeling of Facial Expression and Emotion for Human Communication System. Displays 17, pp. 15-25, Elsevier.

Lee, Y, Terzopoulos, D. and Waters, K. (1995). Realistic modeling for facial animation. Proceedings of SIGGRAPH'95, pp.55-62.

Sera, H., Morishima, S. and Terzopoulos, D. (1996). Physics-based muscle model for mouth shape control. Proceedings of Robot and Human Communication '96 (RO-MAN '96), pp. 207-212.

Multimodal Information Fusion for Improved Speaker Detection

Ashutosh Garg
ECE Department and Beckman Institute
University of Illinois
Urbana, IL 61801

Vladimir Pavlović and James M. Rehg
Compaq Computer Corporation
Cambridge Research Lab
Cambridge, MA 02142
{vladimir,rehg}@crl.dec.com

Abstract

Inferring users' actions and intentions forms an integral part of design and development of any human-computer interface. The presence of noisy and ambiguous sensory data makes this problem challenging. In such scenarios multimodal sources of information may be beneficial if a fusion method can be devised that exploits the sources' interactions. We formulate a learning-based framework for temporal fusion of multiple sensors using input–output dynamic Bayesian networks (IODBNs). We apply this framework to the problem of audio/visual speaker detection in an interactive kiosk application using "off-the-shelf" visual and audio sensors (face, skin, texture, mouth motion, and silence detectors). Detection results obtained in this setup demonstrate numerous benefits of our approach. We also discuss implementation issues and propose extensions of this framework to modeling of multimodal dialogs.

1 Introduction

Human-centered user-interfaces based on vision and speech present challenging sensing problems. Multiple sources of information, including high-level application-specific information, must be combined to infer the user's actions and intentions. Statistical modeling techniques play a critical role in the design and analysis of such systems. Dynamic Bayesian network (DBN) models are an attractive choice, as they combine an intuitive graphical representation with efficient algorithms for inference and learning. Previous work has demonstrated the power of these models in fusing video and audio cues with contextual information and expert knowledge both for speaker detection and other similar applications [3, 6, 4, 2].

Speaker detection is a particularly interesting example of a multi-modal sensing task with application in video conferencing, video indexing and human-computer interaction. Both video and audio sensing provide important information in a multi-person and noisy scenarios. Contextual information or the state of the application is another important component because it often governs the type of interaction. This naturally leads one to consider a special DBN architecture known as input/output DBN (Figure 2(b)). The state of the application forms an input to the system which is, along with the sensory outputs, used to determine the state of the user. We are interested in network models that combine "off-the-shelf" vision and speech sensing with contextual cues.

Estimation of the DBN model parameters is a key step in the design of the detection system. Strengths of the DBN arcs in Figure 1, from context to task variables to sensors, can be automatically learned from data using standard maximum-likelihood (ML) learning schemes, similar to [3]. Sometimes direct application of ML does not suffice. It is then possible to use iteratively improved learning that directly minimizes the error of detecting a speaker. We presented one such approach based on *boosting* in [6].

This paper demonstrates that modeling the contextual states of the application as an input to a DBN together with the learning of continuous sensor distributions can enhance performance of DBNs which fuse temporal data from weak multimodal sensors. We present these results in the context of a network architecture of Figure 1 which infers the state of the speaker who actively interacts with the Genie Casino game. Our evaluation of the learned DBN model indicates its superiority over previous static [8] and dynamic [3, 6] detection models. We also suggest enhancements of this model to include sensor reliability and propose its use in multimodal dialog modeling.

2 Speaker Detection

An estimate of the persons state (whether s/he is or isn't a speaker) is important for the reliable functioning of any speech-based interface. We argue that for a person to be an active speaker, s/he must be expected to speak, face the computer system and actually speak. Visual cues can be useful in deciding whether the person is facing the system and whether he is moving his lips. However, they are not capable on their own to distinguish an active user from an active listener (listener may be smiling or nodding). Audio cues, on the other hand, can detect the presence of relevant audio in the application. Unfortunately, simple audio cues are not sufficient to discriminate a user in front of the system speaking to the system from the same user speaking to another individual. Finally, contextual information describing the "state of the world" also has bearing on when a user is actively speaking. For instance, in certain contexts the user may not be expected to speak at all. Hence, audio and visual cues as well as the context need to be used jointly to infer the active speaker.

We have analyzed the problem of speaker detection in a specific scenario of the Genie Casino Kiosk. The Smart Kiosk [7] developed at Compaq's Cambridge Research Lab (CRL) provides an interface which allows the user to interact with the system using spoken commands. This version of kiosk simulates a multiplayer blackjack game (see Figure 4 for a screen capture.) The user uses a set of spoken commands to interact with the dealer (kiosk) and play the game. The kiosk has a camera mounted on the top that provides visual feedback. A microphone is used to acquire speech input from the user.

We use a set of five "off-the-shelf" visual and audio sensors: the CMU face detector [9], a Gaussian skin color detector [13], a face texture detector, a mouth motion detector, and an audio silence detector. A detailed description of these detectors can be found in [8]. Contextual input provides the state of the application (the blackjack game) which may help in inferring the state of the user.

2.1 Bayesian networks for speaker detection with continuous sensors and contextual input

We adopt a modular approach towards the design of the Bayesian network for speaker detection. We have designed modules for vision and audio tasks separately which are then integrated along with the higher level information.

The graph in Figure 1 shows the vision network for this task. This network takes the output of the vision sensors and outputs the query variables corresponding to visibility and the frontal information of the user. Face detector gives a binary output whereas the output of the texture detector is modeled as a conditional Gaussian distribution whose parameters are *learned* from the training data. We contrast this to the cases studied in [3, 6] where all sensors had binary outputs.

The audio network combines the output of the silence detector and the mouth motion detector. These detectors provide continuous valued output as the measure of the silence and mouth motion respectively. The audio network selected for this task is shown in Figure 1. The output of the audio network corresponds to the probability that the audio in the application corresponds to the user present.

Once constructed, the audio and visual networks are fused to obtain the integrated audio–visual network. The contextual information acts as an input to the model with the sensory observations as the output. The state of the user (e.g. speaker vs. nonspeaker) forms the state of the model and needs to be inferred given the observations and the inputs. The final network obtained is shown in Figure 1.

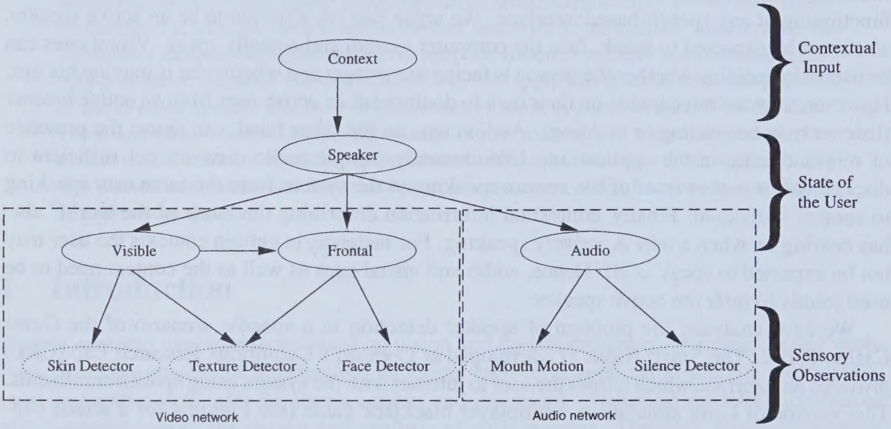
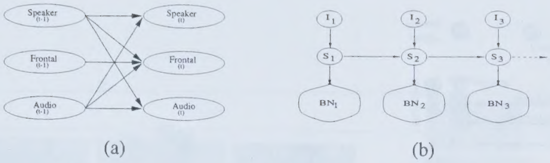


Figure 1: Integrated audio-visual network.

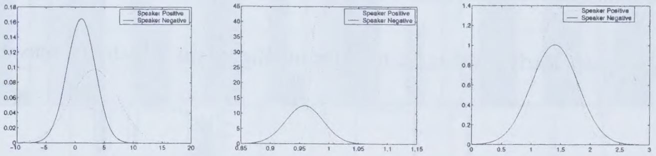
The final step in designing the topology of the speaker detection network involves its temporal aspect. Measurement information from several consecutive time steps can be fused to make a better informed decision. This expert knowledge becomes a part of the speaker detection network once the temporal dependency shown in Figure 2(a) is imposed. Incorporating all of the above elements into a single structure lead to the input/output DBN shown in Figure 2(b). The input/output DBN structure is a generalization of the input/output HMM [1]—here the probabilistic dependencies between the variables are governed by the BN shown in Figure 1. The speaker node is the final speaker detection query node.

The use of continuous valued sensor outputs allows the network to automatically learn optimal sensor models and, in turn, optimal decision thresholds. We model all continuous sensory outputs are as conditional Gaussian distributions, as shown in Figure 3. The learned distributions allow soft sensory decisions which can be superior to discrete sensory outputs in noisy environments.



(a) (b)

Figure 2: (a) Temporal dependencies between the speaker, audio, and frontal nodes at two consecutive time instances. (b) The input output DBN, with contextual information as input and the sensory observations as outputs.



(a) (b) (c)

Figure 3: Learned continuous sensor distributions: (a) silence, (b) skin texture, and (c) mouth motion.

3 Learning dynamic Bayesian networks

Dynamic Bayesian networks are a class of Bayesian networks specifically tailored to model temporal consistency present in data. In addition to describing dependencies among different static variables DBNs describe probabilistic dependencies among variables at different time instances. A set of random variables at each time instance t is represented as a static BN. Out of all the variables in this set temporal dependency is imposed on some. Thanks to its constrained topology efficient inference and learning algorithms, such as forward-backward propagation and Baum-Welch, can be employed in DBNs (see [6] for more details.)

However, classical DBN learning algorithms assume that the selected generative model accurately represents the data. This is often not the case as the selected model is only an approximation of the true process. Recently, Schapire et al. [10] have proposed a method called *boosting* aimed at improving the performance of any simple (classification) model. In particular, their *Adaboost* algorithm "boosts" the classification on a set of data points by linearly combining a number of weak models, each of which is trained to correct "mistakes" of the previous one. We formulated the *Error Feedback DBN* framework for both discrete and continuous sensory models in [6] and [5].

4 Experiments and Results

For convenience, we repeat experimental description presented in [5]. A set of three experiments was conducted using a common data set of five sequences of a user playing the blackjack game in the Genie Casino Kiosk setup. The experimental setup is depicted in Figure 4. The same figure shows some of the recorded frames from the video sequence. Each sequence included audio and video tracks recorded through a camcorder along with frequency encoded contextual information (see Figure 4.) The visual and audio sensors were then applied to audio and video streams. Examples of individual sensor observations (e.g., frontal v.s. non frontal, silence v.s. non silence, etc.) are shown in Figure 5. Abundance of noise and ambiguity in

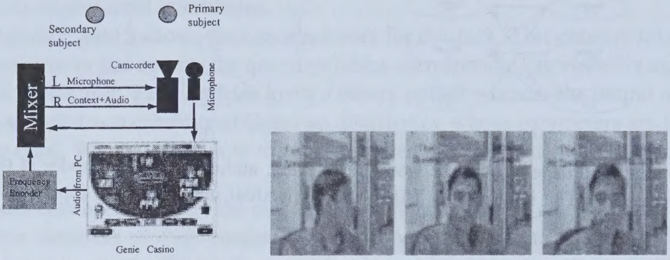


Figure 4: Data collection setup for Genie Casino kiosk and three frames from a test video sequence.

these sensory outputs clearly justifies the need for intelligent yet data-driven sensor fusion.

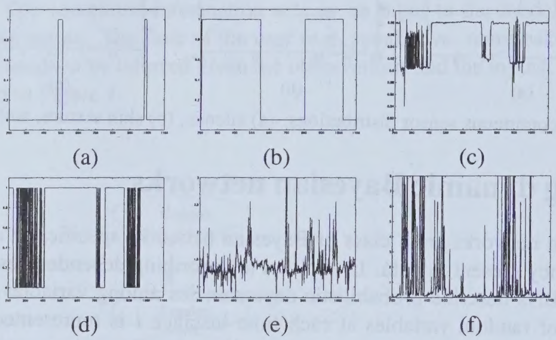


Figure 5: (a) Ground truth for the speaker state: 1 indicates the presence of the speaker. (b) Contextual information: 1 indicates user's turn to play. (c),(d),(e),(f) Outputs of texture, face, mouth motion and silence detectors, respectively.

4.1 Static Bayesian network

The first experiment was done using the static BN of Figure 1 to form the baseline for comparison with the dynamic model. In this experiment all samples of each sequence was considered to be independent of any other sample. Part of the whole data set was considered as the training data and rest was retained for testing. During the training phase, output of the sensors along with the hand labeled values for the hidden nodes (speaker, frontal and audio) were presented to the network.

During testing only the sensor outputs were presented and inference was done to obtain the values for the hidden nodes. Mismatch in any of the three (speaker, frontal, audio) is considered to be an error. An relatively low average accuracy of 77% is obtained (see Figure 6 for results on individual sequences.) The sensor data (as shown in Figure 5) is noisy and it is hard to infer the speaker without making substantial errors. Figure 7(a) shows the ground truth sequence for the state of the speaker and (b) shows the decoded sequence using static BN. However, this detection accuracy is superior to 68% obtained in [3, 6] when discrete sensors and non-input context were used.

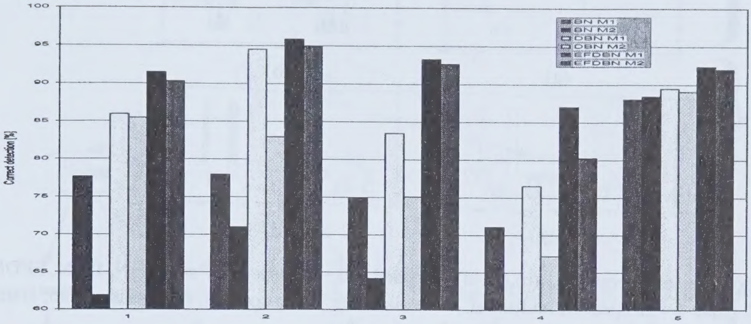


Figure 6: A comparison between the results obtained using static BN, DBN, and EFDBN. M1 and M2 denote the models with continuous and discrete sensory inputs, respectively.

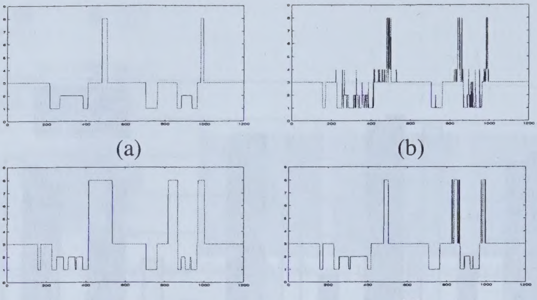
4.2 Input/output DBN

Second experiment was conducted using the input/output DBN model. The standard ML learning algorithm described in Section 3 was employed to learn the dynamic transitional probabilities among frontal, speaker, and audio states. During testing phase a temporal sequence of sensor values was presented to the model and Viterbi decoding was used to find the most likely sequence of the speaker states. Overall, we obtained the accuracy of the speaker detection of 85%, an improvement of $\approx$ 10% over the static BN model. An indicative of this can be seen in actual decoded sequences in Figure 7. The improved performance by the use of DBN stems from the inherent temporal correlation present between the features. Again, the use of continuous sensors and input context produced significant improvement compared to 80% rate of [3, 6].

4.3 Error-feedback DBN

Our final experiment employed the newly designed EFDBN framework for continuous sensors and contextual application input. The learning algorithm described in [5] was used. For a training sequence, we used EFDBN to estimate the parameters which minimized the classification error. A leave-one-out crossvalidation resulted in the overall accuracy of 92%. Figure 6 summarizes classification results on individual sequences. We see that for all the sequences, an improvement of $5 - 10\%$ over the best DBN result is obtained. While the improvement is less dramatic over the 90% detection rate of the EFDBN with discrete sensors and contextual measurement [6], it still remains significant.

The DBN model learned using the EFDBN framework was also applied to the prediction of hidden states. An overall accuracy of 88% was obtained. This indicates, together with the previously noted results, that EFDBN significantly improves the performance of simple DBN classifiers. In comparison of the results with the ones reported in [3, 6], we observe that significant improvement in performance is obtained. This can be attributed to the use of continuous sensory observations and input/output DBN structure.



(a) (b)

(c) (d)

Figure 7: (a) True state sequence. (b),(c),(d) Decoded state sequences by static BN, DBN, EFDBN, respectively. (state 1 - no speaker, no frontal, no audio; state 2 - no speaker, no frontal, audio; state 3 - no speaker, frontal, no audio; state 8 - speaker, frontal, audio)

5 Future work

Input-output DBN framework presented here represents only a first step in building a fully learning-based human–computer interaction system. We now address the issue of sensor reliability which will become important for practical implantation of a speaker detection system. We also propose an alternative use of IODBN framework in multimodal dialog modeling.

5.1 Sensor Reliability

Reliability and confidence of sensors during inference is one crucial aspect of sensor fusion tasks which is not addressed in the original framework. Sensor reliability determines which one of multiple sensor outputs can be reliably used for inferring the speaker state. Reliability can often be modeled as a function of a measurement space that is not observed by the sensor. For instance, the number of skin colored pixels in the whole image can be used as a measure of the reliability of the skin sensor which only observes an image subwindow. A large number of these "outside" pixels may indicate low sensor reliability.

Probabilistic models of sensor reliability can be seamlessly merged with our IODBN framework. A simple way of including reliability of sensors in shown in Figure 8. Similar model has been proposed for face detection in [12] An intuitive reason for this reliability model is that, given the structure in Figure 8, each hidden state-to-sensor pdf now also becomes conditioned on an estimate of sensor's confidence. Different levels of sensor confidence imply different influences of sensor outputs on their conditioning states. In the example above, low skin color confidence could be associated with an almost uniform skin color-to-visible state distribution. Of course, as in the case of our original IODBN framework, these probabilistic rules would be learned from data (see Section 3.)

5.2 Multimodal Dialog Modeling

Another interesting application of the IODBN framework arises once the framework becomes a part of a larger dialog system. Current dialog systems are mostly hand-tuned in the way a dialog is conducted. For example, such systems often assume a predefined dialog flow with either system or mixed initiative. Recently, probabilistic approaches such as [11] have been

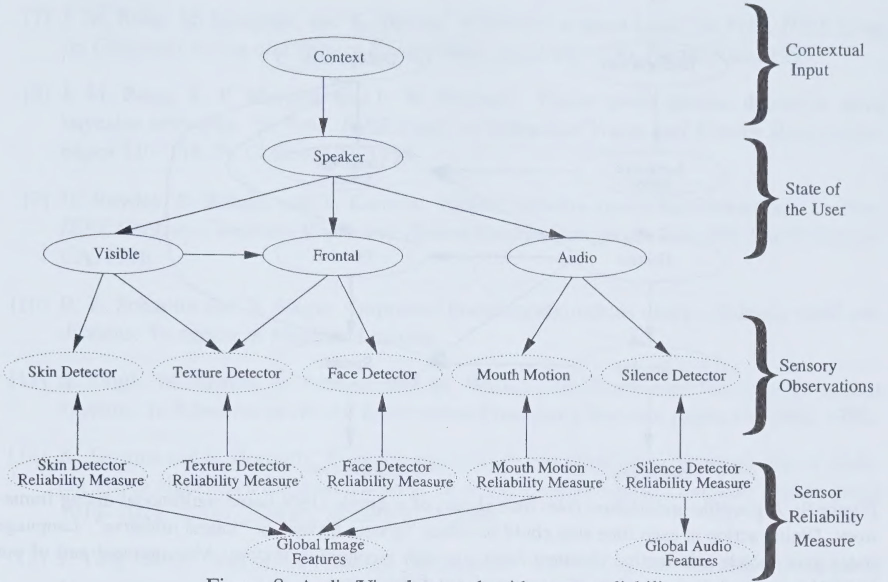

Context
Speaker
Visible
Frontal
Audio
Skin Detector
Texture Detector
Face Detector
Mouth Motion
Silence Detector
Skin Detector Reliability Measure
Texture Detector Reliability Measure
Face Detector Reliability Measure
Mouth Motion Reliability Measure
Silence Detector Reliability Measure
Global Image Features
Global Audio Features
Contextual Input
State of the User
Sensory Observations
Sensor Reliability Measure

Figure 8: Audio/Visual network with sensor reliability.

proposed which offer an optimal choice of initiative based on a current "state of the world" and learned dialog strategies.

Currently, an often ambiguous "state of the world" is solely determined by the information content of an audio stream. Inclusion of information from other modalities, for instance as produced by our IODBN framework, can certainly help in disambiguating the state of the dialog. Most probabilistic dialog systems are modeled as Markov decision processes (MDPs) or partially observable MDPs (POMDPs), which can in turn be represented in a DBN framework. Including our IODBN to that dialog model can now be viewed as augmenting the "state of the world" with another ("speaker") state. An example of such combined framework is depicted in Figure 9. We are currently working on experimental implementations of such dialog systems in our Smart Kiosk setup.

6 Conclusions

We have presented a general purpose framework for learning input/output DBN models for fusion of continuous sensory output and contextual, application-specific input. The framework encompasses an error-feedback learning procedure which can circumvent the effects of simple models and complex data. Significant improvement in classification results obtained for the difficult problem of speaker detection indicate the utility of our framework. We have also demonstrated a general purpose approach to solving man-machine interaction tasks in which DBNs are used to fuse the outputs of simple audio and visual sensors while exploiting their temporal correlation.

We have proposed an augmentation of the IODBN framework to include important information on sensor reliability. We also suggested how a multimodal speaker detection framework

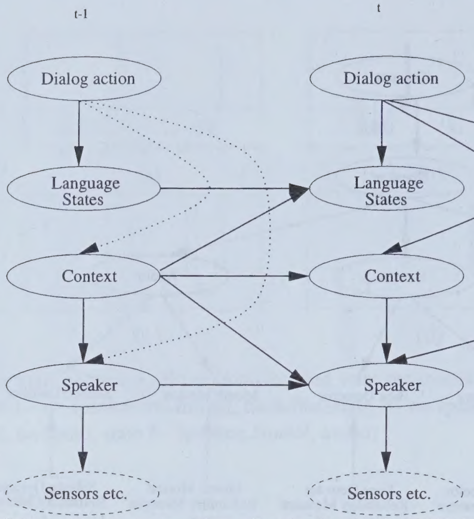


Figure 9: A possible architecture (two time slices) of a simple DBN-based multimodal dialog framework. Dialog action at each time step could be either "system driven" or "mixed initiative". Language states may encode information obtained from a speech recognition system. Measurement part of our IODBN framework is depicted as "Sensors etc." for convenience.

can become an integral part of a complex multimodal dialog-based system.

References

[1] Y. Bengio and P. Frasconi. An input-output HMM architecture. In *Advances in Neural Information Processing Systems 7*, pages 427–434. MIT Press, Cambridge, MA, 1995.

[2] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 994–999, San Juan, PR, 1997.

[3] A. Garg, V. Pavlovic, J. Rehg, and T. S. Huang. Audio–visual speaker detection using dynamic Bayesian networks. In *Proc. of 4rd Intl Conf. Automatic Face and Gesture Rec.*, pages 374–471, Grenbole, France, 2000.

[4] S. Intille and A. Bobick. Representation and visual recognition of complex, multi-agent actions using belief networks. Technical Report 454, MIT Media Lab, Cambridge, MA, 1998.

[5] V. Pavlovic, A. Garg, and J. M. Rehg. Multimodal speaker detection using input/output dynamic bayesian networks. to appear in Third Int'l Conference on Multimodal Interfaces, Beijing, China, October 2000.

[6] V. Pavlovic, A. Garg, and J. M. Rehg. Multimodal speaker detection using error feedback dynamic bayesian networks. In *Computer Vision and Pattern Recognition*, pages 34–41, Hilton Head Island, SC, June 13-15 2000.

[7] J. M. Rehg, M. Loughlin, and K. Waters. Vision for a smart kiosk. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 690–696, Puerto Rico, 1997.

[8] J. M. Rehg, K. P. Murphy, and P. W. Fieguth. Vision-based speaker detection using bayesian networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 110–116, Ft. Collins, CO, 1999.

[9] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 203–208, San Francisco, CA, 1996.

[10] R. E. Schapire and Y. Singer. Improved boosting algorithms using cofidence rated predictions. To appear in Machine Learning.

[11] S. Singh, M. Kearns, D. Litman, and M. Walker. Reinforcement learning for spoken systems. In *Advances in Neural Information Processing Systems*, pages 957–962, 1999.

[12] K. Toyama and E. Horovitz. Bayesian modality fusion: Probabilistic integration of multiple vision algorithms for head tracking. In *Asian Conference on Computer Vision*, pages 1014–1021, Taipei, Taiwan, January 8-11 2000.

[13] J. Yang and A. Waibel. A real-time face tracker. In *Proc. of 3rd Workshop on Appl. of Comp. Vision*, pages 142–147, Sarasota, FL, 1996.

Simple Layout Algorithms To Maintain Network Connectivity Under Faults

Galen H. Sasaki[1]
Department of Electrical, University of Hawaii
2540 Dole Street, Honolulu, HI 96822 USA

Ching-Fong Su and David Blight
Fujitsu Laboratories of America
595 Lawrence Expressway, Sunnyvale, CA 94086-3922 USA

Abstract: A problem of survivable layout of an IP network topology on a wavelength division multiplexed network is considered. Simple layout algorithms are given that have low time complexities. The algorithms are variations of a basic algorithm, that lays out IP links as lightpaths sequentially according to some order and with respect to link costs. Simulations show that the ordering and link costs can significantly affect the performance of the algorithm.

1. Introduction

Wavelength division multiplexing (WDM) is the current technology of choice for broadband backbone networks. WDM long-haul, transport systems are being deployed due to their low cost per bit, good quality of service, and very wide bandwidths. WDM is essentially frequency division multiplexing of optical signals onto optical fibers, and *wavelengths* are equivalent to carrier frequencies (actually, the inverse of carrier frequencies). The technology's capability of providing very wide bandwidths using the existing optical fiber resources make it a good match to meet the exponential growth of Internet traffic.

In this paper we consider the problem of supporting IP networks over WDM networks. Figure 1 shows this two-layer network hierarchy, where the IP links are realized by optical end-to-end connections, referred to as *lightpaths*, at the WDM layer. At the WDM layer are fiber-pairs, each carrying multiple optical signals at different wavelengths. A fiber-pair forms a full-duplex physical link, with the fibers carrying signals in opposite directions. The fiber-pair carries full-duplex optical channels at each wavelength. In this paper, the fiber-pairs will be referred to as *fiber-links*. The fiber-links are terminated by transport equipment that multiplexes/demultiplexes the optical signals so that the fiber-link operates as a collection of parallel, full-duplex optical channels at different wavelengths. Today, lightpaths and WDM channels are 2.5 Gbps (OC-48) and 10 Gbps (OC-192), while 40 Gbps (OC-768) is the emerging technology.

The WDM optical channels from the transport equipment may be cross-connected at *optical cross-connects* (OXCs). The transport equipment has O-E-Os (optical-electrical-optical) equipment, which implies that any pair of optical channels incident to a common OXC may be cross-connected. (Note that without O-E-Os only channels with the same wavelength may be cross-connected.)

---

[1] The author was supported in part by Fujitsu Laboratories of America, and the National Science Foundation under grant NCR-9612846.

Full-duplex lightpaths are formed at the WDM layer, by cross-connecting WDM channels at OXCs. The lightpaths are terminated at interface ports of high speed IP routers, that are connected to OXCs as shown in Figure 1. To the IP routers, lightpaths are their physical links.

Since the IP network has high capacity, and thus supports many end-users, it is important that it is survivable under faults. At a minimum, the network should remain connected under any possible fault. As long as the IP network remains connected, it can reorganize its routing to provide some minimum level of packet transport service, and its control plane will remain operational.

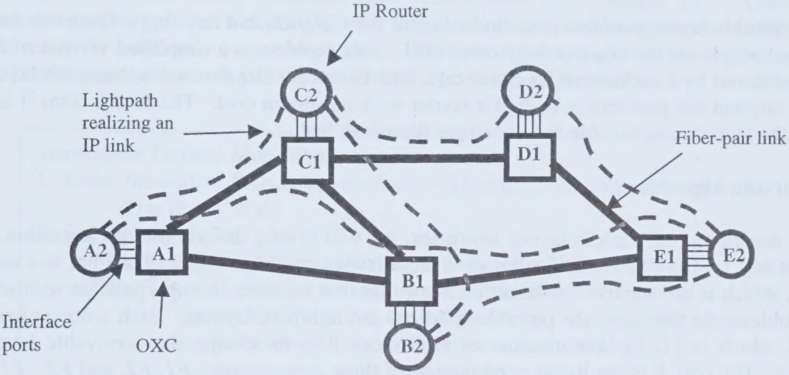


Figure 1. IP network over WDM network.

The type of faults that we consider is the fiber-link failure (e.g., fiber cut). Thus, it is necessary that the IP network topology is *2-connected,* which means that the topology remains connected under any single link failure. Though 2-connectivity is necessary it is not sufficient. If lightpaths (realizing IP links) pass through a common fiber-link then a fiber-link failure will result in multiple IP link failures. This may disconnect the IP network. For example, consider Figure 1. The IP network topology is 2-connected. However, its lightpath layout makes it vulnerable to single fiber-link cuts. In particular, if the fiber-link between OXCs *A1* and *C1* fails then the two IP-links between IP router *A2* and routers *B2* and *C2* fail, leaving IP router *A2* disconnected.

In this paper we consider the problem of laying out the IP network topology over the WDM network so that the former is connected under single fiber-link faults. This was first considered by Crochat and Le Boudec [1], and more recently in [2]. The time complexity of their algorithm for solving the problem can be prohibitive for large networks. We propose simpler algorithms that have lower time complexities.

In Section 2, we describe the network model and survivable layout problem. In Section 3, heuristic layout algorithms with low time complexities are given. Simulation results demonstrating the efficacy are presented in Section 4, and finally concluding remarks are given in Section 5.

2. Network Model and Survivable Layout Problem

There are two network topologies: the IP and WDM network topologies. Both have the same set of nodes V. Each node in V corresponds to an OXC and IP-router pair, where the router is connected to the OXC with interface ports. The fiber-links between the OXCs form the links E_{WDM} of the WDM network. $G_{WDM} = (V, E_{WDM})$ denotes the WDM network topology. The IP network topology is denoted by $G_{IP} = (V, E_{IP})$, where E_{IP} is the collection of IP links. An IP link is realized by a lightpath in the WDM network. A collection of lightpaths that realize the IP links is referred to as a *layout*. We will assume that the WDM network has a large number of wavelengths, so that the number of lightpaths that may traverse a fiber-link is unrestricted.

The *survivable layout problem* is to find a layout for G_{IP} such that any single fiber-link fault in G_{WDM} will leave the G_{IP} topology connected. This problem is a simplified version of the one considered by Crochat and Le Boudec [1, 2]. They consider the case where each layout has a *cost*, and the problem is to find a layout with minimum cost. The cost is zero if and only if the layout is survivable for any single fiber-link fault.

3. Heuristic Algorithms

Before describing our simple layout heuristics, we will briefly discuss the the algorithm of Crochat and Le Boudec [1, 2] for the survivable layout problem. Their algorithm is a *tabu search*, which is an iterative optimization algorithm that searches through possible solutions to a problem. In this case, the possible solutions are lightpath layouts. Each solution has a cost F, which in [1, 2] is a measure of how close it is to solving the survivable layout problem. The cost F is the linear combination of three components: $F1$, $F2$, and $F3$. $F1$ is the sum of fiber-link costs, where the cost of a link e is $\max(0, r(e) - C(e))^2$, $C(e)$ is the capacity of link e, and $r(e)$ is the number of lightpaths on e. (In our case, $C(e)$ is essentially infinite since we assume a very large number of wavelengths.) This cost penalizes fiber-links with load exceeding their link capacities. $F2$ is a measure of the number of IP links broken for a single link fault. $F3$ is the sum of fiber-link costs, where the cost of a link e is $\max(0, t(e) - r(e))^2$ and $t(e)$ is some threshold. The threshold $t(e)$ can be set to 1 to balance lightpath loads across fiber-links.

During one of its iterations, tabu search determines a new layout L_{new} from its current one $L_{current}$. It does this by first creating a collection of layouts that are "perturbations" of $L_{current}$. A perturbation is basically $L_{current}$ with exactly one of its lightpaths randomly rerouted. Then L_{new} is chosen among the perturbations with minimum. Note that each iteration requires $|E_{IP}|$ perturbations, and each perturbation must have its cost computed. Crochat and Le Boudec [1, 2] reported that each iteration takes $O(|E_{WDM}||E_{IP}||V|^2)$ time. The time complexity of an iteration is dominated by the time it takes to evaluate the cost of a candidate layout and in particular cost $F2$.

Thus, the algorithm does not scale well for large networks, especially if it requires many iterations to find a good solution. To address the scalability problem, we investigate simpler layout algorithms that have time complexities on the order of $O(|E_{IP}||V|^2)$. The complexity is dominated by computing $|E_{IP}|$ lightpath routes, each requiring $O(|V|^2)$ shortest path computation.

The algorithms we consider find lightpaths for subsets of IP links at a time. Once a lightpath is computed for an IP link, the IP link is not considered again. The computation of the lightpaths for a subset of size k has time complexity $O(k\ |V|^2)$, which is dominated by k shortest path computations. Thus, the overall time complexity of the algorithms are $O(|E_{IP}|\ |V|^2)$.

Figure 2 shows the basic algorithm. The nodes are considered in some *order*. For each node v, IP links incident to the node, which we denote by $L(v)$, will have their lightpaths computed. Only those IP links that do not yet have lightpaths will have their lightpaths computed, i.e., an IP link will have its lightpath computed only once. Finding lightpaths for the IP links in $L(v)$ can be done by either routing them one-by-one using shortest path routing, or together using *minimum cost link-disjoint path routing* [3]. One-by-one computations are referred to as *single route* computations, while finding the routes together is referred to as a *multiple route* computation. Multiple route computation is implemented by the link-disjoint path routing of [3] and is described in Appendix A.

```
Survivable Layout Algorithm
1. Order the nodes. Denote the nodes by $v(1), v(2),..., v(n)$, where $n = |V|$ and
       $v(k)$ is the k^th node.
2. For each $k = 1, 2, ..., n$,
       2a. Let $L(v(k))$ be the collection of IP links incident to $v(k)$.
       2b. Layout lightpaths for the links in $L(v(k))$ that do not have
               lightpaths. (Note that the layout can be done by single route
               computations or by a multiple route computation.)
```

Figure 2. Basic algorithm to compute lightpath layout of an IP network topology.

Step 2b in Figure 2 is the step for single or multiple lightpath route computations. In the case of single route computation, a lightpath is computed by first having *costs* assigned to fiber-links of G_{WDM}. Then a shortest path computation is applied to find the lightpath. In the case of multiple route computation, first costs are assigned to fiber-links of G_{WDM}. Then minimum cost, link disjoint paths are computed (as described in Appendix A).

By appropriately choosing the costs, one can control the type of paths that will be found. In the case of single route computations, the fiber-links have the following costs.

- *Unit link (UNIT) cost:* Each fiber-link has unit cost. This leads to shortest hop lightpaths, which minimize the amount of bandwidth use. However, this may also lead to an imbalance of lightpath loads on fiber-links. Such an imbalance should be avoided because a fault along a heavily loaded fiber-link will result in faults on many IP links.

- *Quadratic load (QUAD) cost:* Each fiber-link e has cost $(l_e + 1)^2$, where l_e is the number of lightpaths using e. This tends to distribute lightpaths evenly over fiber-links, reducing multiple IP link failures per fiber-link failure. This is similar to cost *F3* of Crochat and Le Boudec [1, 2].

- *Existing Lightpaths-Single Route Computation (EL-S) cost:* This cost depends on the lightpath being computed. The fiber-link costs are as follows. Suppose that s and t are the terminating nodes of the lightpath being computed. A fiber-link incident to s (resp., t) will have high cost K (= $2\,|E_{IP}|\,|V|$) if it carries a lightpath that terminates at s (resp., t). The other links have unit cost. This tends to keep lightpaths that terminate at a common node from using the same links.

  As an example, consider Figure 1 with its existing collection of lightpaths. Suppose a new lightpath is to be computed between routers at *A2* and *E2*. Then the fiber-links *(A1,C1)*, *(B1,E1)*, and *(D1,E1)* will have high cost K because they are incident to OXC *A1* or *E1*, and they carry lightpaths terminating at those OXCs.

In the case of multiple route computations, we use the following fiber-link cost.

- *Existing Lightpaths-Multiple Route Computation (EL-M) cost:* This cost depends on node $v(k)$ of Step 2 of the algorithm (see Figure 2). A fiber-link incident to $v(k)$ that has a lightpath terminating at $v(k)$ has high cost $K = 2\,|E_{IP}|\,|V|$ (same K, as in the *EL-S* cost). The other fiber-links have unit cost. Like the *EL-S* cost, this tends to keep lightpaths that terminate at a common node from using the same links.

  As an example, consider Figure 1 with its existing collection of lightpaths. Suppose new lightpaths were to be computed from the router *A2*. Since the existing lightpaths that terminate at *A2* use fiber-link *(A1,C1)*, the link has cost K and the other links have unit cost.

We should note that with multiple route computation, the survivable layout algorithm may end up with some IP links without lightpaths. This can occur if there are many IP links, but not many fiber-link disjoint paths out of nodes. To take care of the IP links without lightpaths, the survivable lightpath layout algorithm is run again for these links but now for single route computations and *EL-S* cost.

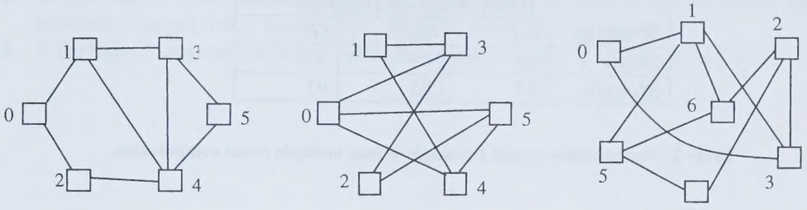
We consider two types of node orderings.

- *Fixed ordering (FIXED):* The ordering is arbitrary and fixed. In this paper, we order according to node IDs.

- *Smallest degree first (DEGR):* This is an ordering according to node degrees, where the degree of a node is the number of IP links incident to it. Presumably nodes with smaller degrees are more likely to be disconnected from a fault, so they are considered first.

4. Performance Results

To compare the options of the survivable lightpath layout algorithm in the previous section, we ran simulations. For each option we ran simulations on three topologies of G_{WDM} shown in Figure 3. Topology *G1* was created by hand, while topologies *G2* and *G3* were randomly created. The random topologies were generated iteratively starting from initial topologies, that have no fiber-links. At each iteration, a fiber-link is added between a random pair of nodes, each pair being equally likely until the topology is 2-connected. However, we disallow multiple links between nodes.



Topology *G1* Topology *G2* Topology *G3*

Figure 3. Three WDM network topologies.

For each WDM topology, 100 random IP topologies (G_{IP}) were created. These topologies were created iteratively in the same way topologies *G2* and *G3* were created. Each IP topology has a layout computed for the WDM topology using the survivable layout algorithm. Table 1 shows the results for single route computations when the node ordering is *FIXED* or *DEGR*, and the fiber-link costs are *UNIT*, *QUAD*, or *EL-S*. There are results for each WDM topology. Each table entry has the number of IP topologies that result in survivable layouts. For example, for WDM topology *G1*, the number of IP topologies that result in survivable layouts is 43 if the node order is *FIXED* and the link cost is *QUAD*.

	WDM Network Topology								
	G1			***G2***			***G3***		
	Link Cost			**Link Cost**			**Link Cost**		
Order	*UNIT*	*QUAD*	*EL-S*	*UNIT*	*QUAD*	*EL-S*	*UNIT*	*QUAD*	*EL-S*
FIXED	33	43	70	33	50	48	34	75	84
DEGR	33	60	84	33	69	69	34	90	92

Table 1. Performance results for single route computation.

The table shows that node ordering and link costs can significantly improve performance. In the case of node ordering, *DEGR* performs at least as well as *FIXED*, and in most cases the performance difference is significant. In the case of link costs, *EL-S* performs best. Thus, the combination of options *DEGR* and *EL-S* works best for single route computation.

Table 2 compares single and multiple route computation. The data for the single route computation for the options *DEGR and EL-S*, which we have shown work best. The table shows that multiple route computations performs about the same or better.

	WDM Network Topology		
Routing	*G1*	*G2*	*G3*
Single	84	69	92
Multiple	84	85	91

Table 2. Performance results for single versus multiple route computation.

5. Concluding Remarks

We propose simple algorithms to compute a survivable lightpath layout for an IP network topology. It is an iterative approach that computes lightpaths in sequence. Link cost and node ordering options were compared for single route computation. It was shown that the choice of node ordering and link costs can result in significant improvement. In the case of node ordering, it is better to first layout lightpaths of nodes with a small number of IP links since they are more vulnerable to faults. In the case of link costs, *UNIT* is the worse since it will not spread lightpaths over links. *QUAD* is better since the link costs tend to balance lightpath load over links. *EL-S* is a little better than *QUAD*, perhaps because its link costs tend to spread lightpath load and it is also specific to the lightpath being routed. We finally compared single and multiple route computation, and found that multiple route computation is about the same or better.

A. Link disjoint path algorithm

In this appendix we discuss how multiple route computations are done for IP links *L(v)*, which are the IP links incident to node *v*. The algorithm is basically a minimum cost, link disjoint path algorithm of [3] on an appropriately modified network topology. The algorithm will compute lightpaths for those IP links in *L(v)* that do not have one. Let *n* denote the number of such IP links, and denote the links by $L'(v) = \{ (v, u_1), (v, u_2), ..., (v, u_n) \}$.

The algorithm of [3] computes minimum cost, link disjoint paths in a graph between a pair of nodes. If the number of disjoint paths is *k* then the algorithm's time complexity is dominated by the time complexity of *k* shortest path computations. Thus, the algorithm's time complexity is $O(k\,|V|^2)$. We apply the link disjoint path algorithm to a modified version of network topology G_{WDM}. In particular, a new (fictitious) node *t* is added. Also added are new links $(u_1, t), (u_2, t), ..., (u_n, t)$, i.e., these are new links connecting the nodes of $\{u_1, u_2, ..., u_n\}$ to node *t*. The new links have zero cost. Then the minimum cost, disjoint path algorithm of [3] is applied to the modified topology G_{WDM}, where the paths terminate at nodes *v* and *t*. Note that the computed paths can be truncated at the nodes $\{u_1, u_2, ..., u_n\}$ to form link disjoint lightpaths for a subset of the IP links in *L'(v)*. Those IP links in *L'(v)* that do not get a lightpath are left for a future lightpath computation.

References

1. O. Crochat, J.-Y. Le Boudec, “Design protection for WDM networks,” *IEEE J. Selected Areas on Communications,* vol. 16, no. 7, pp. 1158-1165, Sept. 1998.
2. O. Crochat, J.-Y. Le Boudec, O. Gerstel, “Protection interoperability for WDM optical networks,” unpublished manuscript. http://lrcwww.epfl.ch/crochat.
3. J. Surballe, “Disjoint paths in a network,” *Networks*, vol. 4, 1977, pp. 125-145.

On Trading Wavelengths with Fibers: A Cost-Performance Based Study*

Arun K. Somani and Mani Mina
Dependable Computing and Networking Laboratory
Electrical and Computer Engineering Department
Iowa State University, Ames, IA 50011-3060
{arun, mmina}@iastate.edu

Abstract

We consider the effect of multiple fibers on WDM networks without wavelength conversion. We study networks with dynamic wavelength routing and develop accurate analytical models to compare various possible options using single- and multiple-fiber networks. We use results of analytical models and simulation-based studies to evaluate the blocking performance and cost of multifiber networks. The number of fibers required providing high performance in multifiber networks and their costs are compared to make a case for using multiple fibers in each link with fewer wavelengths instead of using a single fiber with many wavelengths. Another advantage is that some fault tolerance with respect to a single fiber failure is already built into the system.

1 Introduction

Computer and telecommunication networks are changing the world dramatically. The Internet, mainly based on packet switches, provides very flexible data services such as e-mail and access to the World Wide Web. The Internet is a variable-delay, variable-bandwidth network that provides no guarantee on the quality of service (QoS) in its initial phase. The Internet traffic volume continues to grow exponentially. A conservative estimate of Internet traffic growth is that it is doubling every six months [1]. Various efforts are being made to provide high levels of QoS on packet networks - particularly for voice and other real-time services. With increasing data traffic and high levels of QoS requirement on packet networks, it is desirable to converge the multiple networks around a single packet-based core network [2]. The *Next-Generation Network Internet* (NGI) architecture will converge to share a common high-level architecture as shown in Figure 1a.

The application signals have widely varied characteristics, e.g., signal format and transmission speed. To transport the varied application signals on the optical transport network, a network service layer is needed to map the signals to optical channel signals along with associated "overhead" to assure proper networking functions. This layer captures today's IP and ATM capabilities with statistical multiplexing and QoS guarantee. New protocols such as multiprotocol label switching (MPLS), resource reservation protocol (RSVP), and differentiated services (DiffServ) are likely to play a major role to support the required QoS across a wide set of applications. The network service layer relies entirely on the transport layer for the delivery of multi-gigabit bandwidth where and when it is needed to connect to their peers.

The optical transport layer is capable of delivering multi-gigabit bandwidth with high reliability. It has advanced features such as optical channel routing and switching

---

[*] This research is supported in part by NSF Grant No. ANI9973102 and DARPA Prime Award No. N66001-00-1-8949, Project No. ANI2666, (subcontract 00-S04).

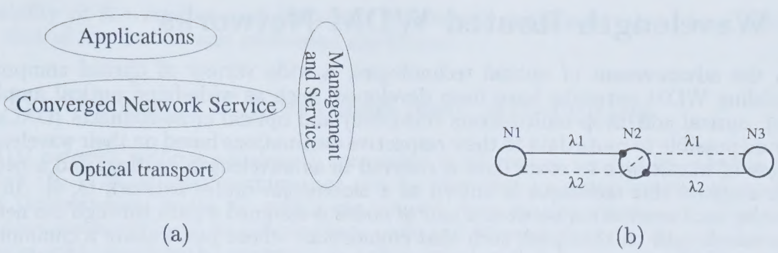


Figure 1: (a) A next-generation network architecture; and (b) Demonstration of wavelength continuity constraint on a two-hop path in a network.

and supports flexible, scalable, and reliable transport of a wide variety of client signals at ultra-high speed. Transport networking enables the service layer to operate more effectively, freeing them from constraints of physical topology to focus on the sufficiently large challenge of meeting service requirements.

Over the past two decades, optical fibers have revolutionized the communications industry. A natural approach to utilize the fiber bandwidth efficiently is to partition the usable bandwidth into non-overlapping wavelength bands. Each wavelength, operating at several gigabits per second, is used at the electronic speed of the end-users. This mechanism is called wavelength division multiplexing (WDM) [3], and is the most promising candidate for improving fiber bandwidth utilization in future optical networks. The research, development, and deployment of WDM technology are now evolving at a rapid pace to fulfill the increasing bandwidth requirement and deploy new network services.

The promise of broadband communication capability in gigahertz and beyond has been a great driving force for accelerated development of WDM and DWDM networks. In 1980's, Deutche Telecom introduced the technology into their BIUFON MANs. That was a 10-channel system (with 36-nm channel spacing and 20-nm signal pass bandwidths). The system was assembled with parts from AEG (today is a part of Daimler-Benz), Amphenol, and Corning France [4].

During the second half of the last decade some noteworthy implementations were commercialized and multiple units were sold internationally, such as Pirelli's 4-wavelength, AT&T's 8-wavelength, and IBM's 20-wavelength (IBM's 9729) systems. In order to achieve more bandwidth out of the existing fibers few DWDM systems were proposed as commercial possibility in 1998-1999. Eventually in 1999, Lucent Technologies won the first order for an 80-channel system, and Pirelli offered a system that is scalable up to 128 channels. In all of these cases the contracts were for systems that are able to be upgraded in the future.

While technologically feasible, such complete systems with full functionality are a huge financial burden on the customer. Consequently, companies implement systems that are affordable and offer real possibility of future expansion. Due to cost factors in 1999-2000 only 4 to 8 of the 80 to 128 channels are being utilized. Some companies are contracting for custom-made solutions up to 32 channels [5]. Meanwhile few companies with established SONET product lines have announced platforms that are scalable up to 32 wavelengths (i.e. Alidian's OSN 4800 and Alcatel's Optinex 1686WM).

The opto-electronic components cost is one of the factors that has hindered the fast penetrations of optical networks in the market. Developers keep looking for new fabrication methods to reduce the prices of switches, transmitters, and receivers. It is difficult to predict when the more economical methods, as well as economy of scale factors, will bring the optical components to a reasonable price level compared to the digital counter parts. It should be noted that these commercial products are sometimes a few orders of magnitude more expensive than the electronic counterparts.

2 Wavelength-Routed WDM Networks

With the advancement of optical technologies, a wide variety of optical components of building WDM networks have been developed, such as wide-band optical amplifiers (OAs), optical add/drop multiplexers (OADMs) and optical cross-connects (OXCs). It becomes possible to route data to their respective destinations based on their wavelengths. The use of wavelength to route data is referred to as wavelength routing, and a network which employs this technique is known as a *wavelength-routed* network [3, 6]. In such networks, each connection between a pair of nodes is assigned a path through the network and a wavelength on that path such that connections whose paths share a common link in the network are assigned different wavelengths. The optical communication path between two nodes is called a *lightpath.* All-optical networks employing wavelength-division multiplexing and wavelength routing are a viable solution for future wide-area networks (WANs) and metropolitan-area networks (MANs). These wavelength-routed WDM networks offer protocol transparency and simplified management and processing compared to routing in telecommunications systems using digital cross-connects [3, 6].

In a wavelength-routed WDM network, the path of a signal is determined by the location of the signal transmitter, the wavelength on which it is transmitted, and the state of the network devices. An example of such a network with 2 wavelengths on each link is shown in Figure 1b. There are two sessions that are in progress, one from node 1 to node 2 using wavelength λ_1, another from node 2 to node 3 using wavelength λ_2. A connection request from node 1 to node 3 has to be blocked, although free wavelengths are available on both link 1 and link 2. This is because of the wavelength continuity constraint, that is, the same wavelength must be assigned to a connection on every link on a path if wavelength converters are not available at the switching nodes. Connection requests encounters higher blocking probability than it does in electronic-switched networks because of the wavelength continuity constraint.

2.1 Routing and Wavelength Assignment (RWA) Algorithms

Routing and wavelength assignment algorithms are responsible for selecting a suitable route and a wavelength among the many possible choices for establishing a connection. Good routing and wavelength assignment algorithms are critically important to improving the performance of WDM networks.

The routing algorithms can be broadly classified into two, namely, *static routing* and *dynamic routing.* In the static routing, the routes for node- pairs are fixed, i.e., the routes do not change with the network status. The static routing typically includes *fixed-path routing (FPR)* and *alternate path routing (APR).* In the dynamic routing, the routes for node-pairs are dynamically selected according to the current network status. A typical example of the dynamic routing is least-congestion routing (LCR).

Wavelength assignment problem is a unique problem in WDM networks. Due to Wavelength continuity constraint, the same wavelength has to be free on all of the links of a path for establishing a connection in all-optical WDM networks without wavelength conversion. If full-range wavelength converter is available at every node, wavelength assignment is a trivial problem. However, the technology of all-optical wavelength conversion is not mature yet. Wavelength converters are likely to be costly devices. Therefore, good wavelength assignment algorithms along with routing algorithms are critically important to improving the network performance and reducing the network cost.

Several algorithms have been proposed for wavelength assignment: *Random* that chooses one of the available wavelengths randomly with a uniform distribution; *First-Fit* that assumes that the wavelengths are arbitrarily ordered, and checks the status of the wavelengths sequentially and chooses the first available wavelength to establish a connection; *Most-Used* selects a free wavelength that is already being used on the most number of links in the network; and *Least-Used* that selects the free wavelength that is used on the least number of links in the network. The chosen wavelength is used to establish a connection. The results in [6] show that both of the random and the least-used algorithms distribute the load evenly over the wavelengths. However, the blocking probability of the random and least-used wavelength assignment algorithms is higher than that of the first-fit and most-used algorithms.

3 Wavelength Conversion and Alternates

A WDM network without wavelength conversion is referred as *wavelength selective (WS)* network [7]. The network performance can be improved by using wavelength converters at the switching nodes [8], which can convert data on one wavelength along a link into other wavelengths at an intermediate node and forward it along the next link. The networks with wavelength conversion is called *wavelength interchanging (WI)* networks [7]. The first solution for wavelength conversion is opto-electronic wavelength conversion, in which the optical signal is converted into the electronic domain first. The electronic signal is then used to drive the input of a tunable laser tuned to the desired wavelength of the output. Since this technique is not transparent to data bit rate and data format, which is one of the major advantages of using optical networking, the opto-electronic wavelength conversion is not preferred to using in the future networks. All-optical wavelength converters are being prototyped in research laboratories [8]. Networks with sparse wavelength conversion [9] and limited wavelength conversion [10] have been analyzed in great details. However, the wavelength conversion techniques have not matured yet. Wavelength converters are likely to remain costly devices in the near future.

An alternative solution to conquer the wavelength continuity constraint is to use multifiber WDM networks [11, 12, 13, 14, 15]. In multifiber networks, each link consists of multiple fibers, and each fiber carries information on multiple wavelengths. A wavelength that cannot continue on the next hop can be switched to another fiber using an optical cross-connect (OXC) if the same wavelength is free on one of the other fibers. Moreover, may also help to support the dramatically increasing bandwidth requirement.

We study the blocking performance of multifiber WDM networks with different routing and wavelength assignment algorithms. A parametric system cost model for both multifiber networks and single-fiber wavelength-convertible WDM networks is also developed. By comparing the cost of different network configurations, we show that a multifiber network is a cost-effective solution under current technology.

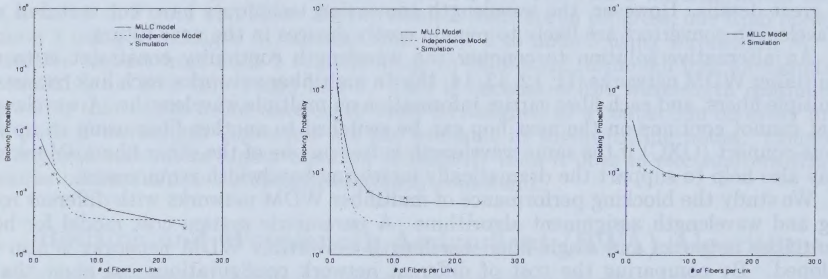
4 Multifiber WDM Networks

Multifiber WDM networks have been studied in [11, 12, 13, 14, 15]. Much research has also been done in the literature on WDM networks with limited wavelength conversion [9, 10]. Assuming Poisson input traffic with arrival rate λ at each node, exponentially distributed call holding time with mean $1/\mu$, a single pre-selected for each source-destination (s-d) pair, a random wavelength assignment, and using analytical and simulation models, it can be shown that a limited number of fibers, each with a fewer wavelength, is sufficient to provide similar performance as that in full-wavelength-convertible networks. Let F denote the number of fibers per link and W denote the number of wavelengths per fiber. We assume that they are the same on all links and fibers, respectively. If the wavelength is not free on all of the F fibers, the request is blocked on this wavelength. No wavelength converter is available at any node.

If we do not account for the cost of wavelength converters, the cost of a multifiber network is likely to be higher than a single-fiber network (more amplifiers and multiplexer/demultiplexer). It should however be noted that a fiber with fewer wavelengths may be able to use cheaper components (like Muxes and Demuxes) for termination. The design goal of a multifiber network is to achieve high performance with the minimum number of fibers per link. An important problem here is to determine how many fibers are required on each link to guarantee performance that is similar to a network with full-range wavelength converters at every node.

Using a multifiber link-load correlation (MLLC) and link-load independent model, we show the results of blocking probability for two regular topologies, the ring and the mesh-torus networks, and an irregular NSF T1 backbone network (NSFnet) [6]. The blocking probability on a l-hop path can be computed recursively by viewing the first $l-1$ hops as the first hop and the lth hop as the second hop of a two-hop path [9]. A detailed analysis has been presented in [14, 15]. We are interested in finding the effect of multifibers on these networks. The question we answer is how many fibers, each with fewer wavelength are required to provide similar performance as that in a full-wavelength-convertible network.

For the results in Figure 2, the link capacity in all the networks has been fixed at 24 light channels, i.e., $FW = 24$ on each link. We vary the number of fibers on each link, F, to be $1, 2, 3, 4, 6, 8, 12$, and 24, and the number of wavelengths on each fiber is varied as $W = 24/F$ accordingly. We observed that 6 fibers per link in the ring and 4 fibers per link in the mesh-torus, are sufficient to provide similar performance as that in full-wavelength-convertible networks (F=24, W=1). We also observe that the full wavelength conversion (F=24, W=1) in the NSFnet does not improve much the performance compared to no wavelength conversion (F=1, W=24). It is also interesting to note that only 4 fibers per link is sufficient to provide similar performance as that of using full wavelength converters in the NSFnet. An important conclusion is that a multifiber network has similar blocking performance as that of a full-wavelength-convertible network, if we select the wavelength-fiber-pairs adequately. A limited number of fibers is sufficient to guarantee high network performance. Thus, multifiber WDM networks without wavelength conversion is not only a feasible, but also a desirable choice under current technologies.



(a) (b) (c)

Figure 2: Blocking probability versus the number of fibers in a (a) 10 node unidirectional ring, (b) 5×5 mesh-torus, and (c) NSFnet network.

One can also use alternate routing mechanisms in multifiber systems. For example, in the alternate-path routing (APR), a set of paths are pre-computed statically for each s-d pair and stored sequentially at the sources nodes according to specified criteria, e.g., the path length. Upon the arrival of a connection request at a source node, the paths are searched sequentially. The first path that has free wavelengths available is selected, and one wavelength is randomly selected to establish the connection. The request is blocked only if all of the candidate paths have no free wavelengths. Alternate-path routing improves the single-fiber network performance significantly. The same [15] can also be used for multifiber networks. We have observed that routing algorithms affect the benefits of using multiple fibers on each link in WDM networks. The number of fibers required to provide high performance using alternate-path routing is slightly higher than that required using fixed-path routing. However, still a limited number of fibers are sufficient to guarantee high performance.

5 Cost Model

In the previous section we introduced multiple fibers on each link as an alternate solution to overcome the wavelength continuity constraint. Using analysis and simulation results, we showed that a limited number of fibers per link are sufficient to guarantee that the network blocking performance is similar to that of networks with full wavelength conversion at every node. However, the cost of a multifiber network cannot be compared to a wavelength-convertible network without having a good cost model. In this section, we develop a parametric cost model for both multifiber networks and single-fiber wavelength-convertible WDM networks. By comparing the cost of different network configurations, we show that a multifiber network is a cost-effective solution with current technology.

5.1 Cost Parameters

We study a general network topology represented by $G(V, E, F, W)$, which shows that the network has $|V|$ nodes and $|E|$ links. We assume that every link has the same number of fibers per link, F, and the same number of wavelengths per fiber, W. We consider three cost factors in a WDM network: the cost related to (1) a cable (C_{Cable} cost), (2) to a fiber (C_{Fiber} cost), and (3) to a switch (C_{Switch} cost).

Cable Cost. The cable cost, C_{Cable}, includes digging cost, leasing cost, right of way cost, and cable maintenance cost, etc., which are required before any capacity can be used on a link. Different link configurations, i.e., the number of fibers per link (F) and the number of wavelengths per fiber (W), do not affect the cable cost significantly. To simplify the cost comparison, we therefore assume that the cable cost of a multifiber network is the same as that of a single-fiber wavelength-convertible network. Another assumption here is that either dark fibers are available ($C_{Cable} = 0$) or new fibers have to be laid out in both multifiber and single-fiber networks.

Fiber Cost. The fiber cost, C_{Fiber}, is the combination of costs associated with fiber in a link. The fiber cost typically consists of the cost of physical fibers, optical amplifiers, dispersion compensation components, multiplexers and demultiplexers to terminate a fiber, and signal regenerators if installed in the network. Since we consider the fiber cost as a combination of the costs associated with the fiber, the number of wavelengths that a fiber carries affects the fiber cost significantly. One reason is that the components handling more wavelengths costs more than the components that handle fewer wavelengths. For example a dispersion compensation component that can handle 80 wavelengths on each fiber may cost several times more than a compensation component that can only compensate the signals on a fiber consisted of 20 wavelengths. Another reason is that the power budget of optical amplifiers such as EDFA decreases with additional wavelengths. Consequently, we either have to increase the laser power or decrease the distance between two optical amplifiers. In either case, the cost of optical amplifiers for each fiber increases with the increasing number of wavelengths per fiber. Therefore the number of wavelengths per fiber, w, has to be considered as a parameter in the fiber cost. We introduce a uniformed parameter, $\bar{C}_{Fiber}(w)$, to describe the fiber cost of a network. $\bar{C}_{Fiber}(w)$ is defined as the cost of a fiber with one-unit length that carries w wavelengths. Let i be a link of the network, $1 \leq i \leq |E|$. $\bar{C}_{Fiber}(w)$ can be computed as

$$\bar{C}_{Fiber}(w) = \frac{\sum_{i=1}^{|E|} C_{Fiber}(i, w)}{\sum_{i=1}^{|E|} L(i)} , \tag{1}$$

where $L(i)$ is the length of link i and $C_{Fiber}(i, w)$ is the cost of fiber with w wavelengths on link i. $C_{Fiber}(i, w)$ is calculated by adding up all the cost associated with one fiber on link i. The total fiber cost of a multifiber network with parameters F and W is given by

$$C_{Fiber}(F, W) = F \times \bar{C}_{Fiber}(W) \times \sum_{i=1}^{|E|} L(i) . \tag{2}$$

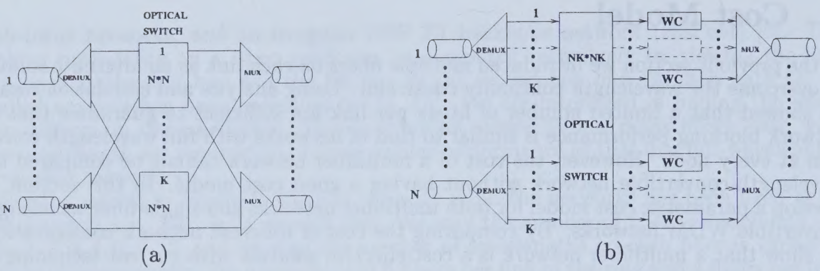


Figure 3: Architecture of nodes (a) without and (b) with wavelength converters

Switch Cost. Many components contribute to the cost of a switch, C_{switch}. To simplify the cost model of an optical switch, we consider two general switch architectures without and with wavelength converters, as shown in Figure 3. Since the number of output ports for Demuxes and input ports for Muxes are same in two switches (for the scenarios we will consider for comparison), we ignore their costs. The actual cost depends on the architecture and technology used in each switch. We account for the cost of switching elements (C_{SE}) in a switch and the cost of wavelength converters (C_{conv}) if incorporated. Dedicated converters are used for each output for each wavelength in the switch as shown in Figure 3b. A converter can also be shared by multiple wavelengths. If so, the cost of a wavelength converter would also be shared.

We assume that the optical switch blocks at each node are non-blocking Batcher-Banyan switches made up of 2×2 switching elements. Let M_{WC} be the number of switching elements required by a node equipped with wavelength converters and N input fibers each with K wavelengths. M_{NWC} denotes the number of switching elements in a node with no wavelength converters and N input fibers each with K wavelengths. M_{WC} and M_{NWC} have been calculated in [7] and are as follows.

$$M_{WC}(N,K) = (NK/4) \times (3 + log_2 NK)(log_2 NK) \tag{3}$$

$$M_{NWC}(N,K) = (NK/4) \times (3 + log_2 N)(log_2 N) \tag{4}$$

Cost Difference. Our goal is to compare networks that have nodes with single fiber input links and full wavelength conversion and multiple fibers input links and no wavelength conversion. To make a fare comparison for the case in hand, we fix the capacity on each link, that is the product of the number of fibers F in each link and the number of wavelengths W carried by each fiber, as constant denoted by D, i.e., $D = FW$. Notice that the number of output ports at Demuxes and the number of input ports at Muxes are same in the two cases and $M_{WC} > M_{NWC}$. Thus a wavelength-convertible optical switch may cost more than a switch without wavelength conversion, not only because of the extra cost of converters, but also the cost of increased number of switching elements.

Let e_j denote the number of links connected to node j. Note the ($\sum_{j=1}^{|V|} e_j = 2|E|$. The switch cost, $C_{Switch-WC}$, for a single-fiber network with $|V|$ nodes, each equipped with wavelength conversion, can be computed by setting $N = e_j$ and $K = D$ for node j. The number of wavelength converters at node j is $D \times e_j$. The cost of consisting of two parts, switching elements and the converters, after ignoring the cost of Demuxes and Muxes is then given by

$$C_{Switch-WC} = C_{SE} \times (\sum_{j=1}^{|V|} M_{WC}(e_j, D)) + C_{conv} \times (\sum_{j=1}^{|V|} De_j) \; . \tag{5}$$

The corresponding in a multifiber network without wavelength conversion is obtained by setting $N = e_j \times F$ (F fibers in each of the e_j links) and $K = D/F$ and is given by

$$C_{Switch-NWC} = C_{SE} \times (\sum_{j=1}^{|V|} M_{NWC}(Fe_j, D/F)) \ . \tag{6}$$

Let C_{diff} be the cost difference between the cost of a single-fiber wavelength-convertible network and the cost of a multifiber network without wavelength conversion. Since, the cost of cable is the same in the two cases, we can divide the total cost difference into two parts: (1) fiber cost difference (C^f_{diff}), and (2) switch cost difference. The switch cost difference again consists of two parts, converters cost difference (C^c_{diff}) and SEs cost difference (C^s_{diff}). Then the cost difference between a single-fiber wavelength-convertible network and a multifiber network, C_{diff}, is given by

$$C_{diff} = C^f_{diff} + C^c_{diff} + C^s_{diff} \ . \tag{7}$$

The three factors with the assumption that $\bar{C}_{Fiber}(D) = \gamma_{DW} \times \bar{C}_{Fiber}(W)$ are as follows.

$$\begin{aligned} C^f_{diff} &= \sum_{i=1}^{|E|}(\bar{C}_{Fiber}(D) - F \times \bar{C}_{Fiber}(D/F)) \times L(i) \\ &= (\bar{C}_{Fiber}(W)) \times (\gamma_{DW} - F) \times (\sum_{i=1}^{|E|} L(i)) \end{aligned} \tag{8}$$

$$C^c_{diff} = C_{conv} \times (\sum_{j=1}^{|V|} De_j) - 0 = 2 \times C_{conv} \times D \times |E| \tag{9}$$

$$\begin{aligned} C^s_{diff} &= C_{SE} \times \sum_{j=1}^{|V|} M_{WC}(e_j, D) - M_{NWC}(Fe_j, D/F) \\ &= C_{SE} \times (\sum_{j=1}^{|V|}(e_jD/4) \times (3 + log_2(e_je_jDF)) \times (log_2(D/F)) \end{aligned} \tag{10}$$

We can clearly see the tradeoff of using multiple fibers in each link or wavelength converters in an optical switch to improve the network performance from the above analysis. Since it can be expected that $\gamma_{DW} < F$ we know from Eq. 8 that $C^f_{diff} < 0$. Eqs. 10 shows that $C^s_{diff} > 0$. And the cost of the converter is of course additional in a single-fiber wavelength-convertible network. To compute the cost difference numerically, we need to unify the cost factors. We choose $\bar{C}_{Fiber}(W)$ and C_{conv} as base cost units.

6 Numerical Results

We studied the blocking performance of the NSFnet in the previous section. We study the cost issue of the NSFnet in this section to see the tradeoff of using multiple fibers per link over single fiber network with wavelength conversion. The NSFnet consists of 14 nodes and 21 links, i.e., $|V| = 14$ and $|E| = 21$. The length of each link, $L(i), i = 1, \ldots, 21$, is shown in Table 1. The link-length is assumed the driving distance between the two nodes. Among the 14 nodes in the NSFnet, 10 nodes h ave input/output degrees $e_j = 3$, 2 nodes have $e_j = 4$, and 2 nodes have $e_j = 2$. In the numerical calculation, for simplicity, we assume that the link capacity is fixed at 32, i.e., $D = 32$. In the multifiber configuration of the NSFnst we studied, we fix the number of fibers per link at 4, i.e., $F = 4$, and

Table 1: Link lengths in miles in the NSFnet

Link	Length	Link	Length	Link	Length
NY - MD	375	NY - MI	643	NY - PA	286
NJ - MI	860	NJ - MD	160	NJ - PA	235
MD - TX	1603	PA - IL	710	PA - GA	883
IL - NE	678	MI - UT	1788	NE - CO	428
GA - TX	1111	CO - TX	813	CO - UT	475
CA2 - CA1	345	TX - CA2	1565	CA2 - WA	986
WA - IL	2031	WA - CA1	1000	UT - CA1	775

correspondingly the number of wavelengths per fiber is $W = 8$. Using these parameters and the node degree for the NSFnet nodes we obtain.

$$\begin{aligned} C^f_{diff} &= (\bar{C}_{Fiber}(W)) \times (\gamma_{DW} - F) \times (\sum_{i=1}^{|E|} L(i)) = 17750 \times (\gamma_{DW} - 4) \times \bar{C}_{Fiber}(8) \\ C^c_{diff} &= C_{conv} \times (\sum_{j=1}^{|V|} De_j) - 0 = 1344 \times C_{conv} \\ C^s_{diff} &= C_{SE} \times (\sum_{j=1}^{|V|} (e_j D/4) \times (3 + log_2(e_j e_j DF)) \times (log_2(D/F)) = 13320 \times C_{SE} \end{aligned}$$

The overall cost difference between the two design is

$$C_{diff} = 17750 \times (\gamma_{DW} - 4) \times \bar{C}_{Fiber}(8) + 1344 \times C_{conv} + 13320 \times C_{SE} \,. \qquad (11)$$

The cost of a unit-length fiber with W wavelengths on it, $\bar{C}_{Fiber}(W)$, depend on many factors, e.g., the technologies used for optical amplifiers and the fibers used that may or may not require dispersion compensation etc. The cost may thus vary significantly in different cases. Similarly, the cost of a wavelength converter, C_{conv}, may also vary depending on what technologies are developed. If we assume that the cost ratio of a fiber with 32 wavelengths to a fiber with 8 wavelengths, γ_{DW}, is 2, and C_{SE} is 100 times cheaper than C_{conv}[1], i.e., $C_{SE} = 0.01 \times C_{conv}$, then we know from Eq. (11) that the overall cost difference between the two network configurations is zero if C_{conv} is about 25 times of $\bar{C}_{Fiber}(8)$. With more expensive converters, the multifiber networks are a better choice.

New technologies of installing fibers in metropolitan-area networks and local-area networks have been developed to reduce the fiber cost. In these networks, optical amplifiers may not be required because of the short distance between any two nodes. Therefore the fiber cost may not play a significant role in the overall network cost in MANs and LANs. Moreover, the technologies for wavelength conversion are not mature yet. The cost of wavelength converters is likely to remain high in the near future. Thus multifiber networks is a cost-effective solution and attractive option for even wide-area networks.

7 Conclusions and Future Work

WDM technology has revolutionized wide-area networks (WAN) by enabling huge increase in the capacity of a single fiber. Wavelength-routed all-optical networks cost-effectively improve the scalability, flexibility, and reliability of not only backbone networks, but also metropolitan-area networks (MAN) and local-access networks (LAN).

---

[1] The converter is not yet available, so it is an estimate.

Different technologies of converting one wavelength to another wavelength have been demonstrated in laboratories. However, none of them is currently available in the market yet. The cost of an all-optical wavelength converter is likely to remain high in the near future. Multi-fiber WDM network is an alternative solution to conquer the wavelength continuity constraint. In multi-fiber WDM networks, a wavelength that cannot continue on the next hop can be switched to another fiber using an optical cross-connect (OXC) if the same wavelength is free on one of the other fibers. We study the effect of multiple fibers in all- optical WDM networks. The results show that a multi-fiber network has similar blocking performance as that of a full-wavelength-convertible network, if we select the wavelength-fiber-pairs adequately. A limited number of fibers are sufficient to guarantee high network performance. We also demonstrate that a multifiber network is a cost-effective solution.

References

[1] M. Shariff, "Packet over SONET Fuels New IP Transport Paradigm", *Lightwave Mag.*, p. 33, April 1998.

[2] Douglas C. Dowden, Richard D. Gitlin, and Robert L. Martin, "Next-Generation Networks", *BELL LABS Technology Journal*, Vol. 3, No. 4, pp. 3-14, Oct- Dec. 1999.

[3] R. Ramaswami and K.N. Sivarajan, *Optical networks: a practical perspective*, Morgan-Kaufman, San Francisco, CA, 1998.

[4] Terry Edwards , "Gigahertz and Teraherts technologies for broadband communication," Artech House, Inc., Copyright 2000, Norwood, MA.

[5] "Australia's first 32-channel DWDM," Altcatel news and events Wednesday, June 7, 2000, www.aleatel,com.au.

[6] D. Banerjee and B. Mukherjee, "Practical approaches for routing and wavelength assignment in all-optical wavelength-routed networks," *IEEE Journal on Selected Areas in Communications*, Vol. 14, pp. 903-908, June 1996.

[7] G. Jeong and E. Aynoglu, "Comparison of Wavelength-Interchanging and Wavelength-Selective Cross-Connects in Multiwavelength All-Optical Networks", *Proc. IEEE INFOCOM '96*, pp. 156–163, 1996.

[8] J. M. Wiesenfeld, "Wavelength conversion techniques," *Proc. OFC '96*, pp. 71–72, 1996.

[9] S. Subramaniam, M. Azizoğlu, and A. K. Somani, "All-Optical Networks with Sparse Wavelength Conversion", *IEEE/ACM Transactions on Networking*, Vol.4, pp.544-557, Aug. 1996.

[10] J. Yates, J. Lacey, D. Everitt, and M. Summerfield, "Limited-range wavelength translation in all-optical networks", *Proc. IEEE INFOCOM '96*, pp. 954- 961, March 1996.

[11] H. Obara, H. Masuda, K. Suzuki, and K. Aida, "Multifiber wavelength-division multiplexed ring network architecture for Tera-bit/s throughput", *Proc. IEEE ICC '98*, Vol. 2, pp. 921-925, June 1998.

[12] S. Baroni, P. Bayvel, R. Gibbens and S. K. Korotky, "Analysis and Design of Resilient Multifiber Wavelength-Routed Optical Transport Networks", *Journal of Lightwave Technology*, Vol. 17, No. 5, pp. 743-758, May 1999.

[13] N. Wauters and P. Demeester, "Wavelength Conversion in Optical Multi-Wavelength Multi-fiber Transport Networks," *Intl.Journal of Optoelectronics*, Vol. 11, No. 1, pp. 53-70, Jan./Feb. 1997.

[14] Ling Li and Arun K. Somani, "Fiber Requirement in Multifiber WDM Networks with Alternate-Path Routing", *Proc. ICCCN '99*, Boston, MA, 1999.

[15] Ling Li and Arun K. Somani, "A New Analytical Model for Multifiber WDM Networks", *Proc. Globecom '99*, pp. 1007 - 1011, Rio de Janeiro, Brazil, December 1999.

Topology Design in WDM Rings to Minimize Electronic Routing: Efficient Computation of Tight Bounds *

Rudra Dutta, George N. Rouskas
Department of Computer Science, North Carolina State University, Raleigh, NC 27695-7534
{rdutta,rouskas}@csc.ncsu.edu

Abstract

We consider the problem of designing a virtual topology to minimize electronic routing, that is, grooming traffic, in wavelength routed optical rings. The full virtual topology design problem is NP-hard even in the restricted case where the physical topology is a ring, and heuristics have been proposed in the literature. We present a new framework of bounds which can be used to evaluate the performance of heuristics, and which requires significantly less computation than evaluating the optimal solution. This framework consists of a sequence of bounds, both upper and lower, in which each successive bound is at least as strong as the previous one. The number of bounds to be evaluated for a given problem instance is only limited by the computational power available. The bounds are based on decomposing the ring into sets of nodes arranged in a path, and adopting the locally optimal topology within each set. While we only consider the objective of minimizing electronic routing in this paper, our approach to obtaining the sequence of bounds can be applied to many virtual topology problems on rings. The upper bounds we obtain also provide a useful series of heuristic solutions.

1 Introduction

In recent years, wavelength routed optical networks have been seen to be an attractive architecture for the next generation of backbone networks. This is due to the high bandwidth in fibers with *wavelength division multiplexing* (WDM) and the ability to trade off some of the bandwidth for elimination of electro-optic processing delays using *wavelength routing* [4]. It has also been noted in literature that, at least in the short term, physical topologies in the forms of rings are of more interest because of available higher layer protocols such as SONET/SDH [6, 5].

Two concerns have recently emerged in this area: it has been recognized that the the cost of network components, specifically electro-optic equipment and SONET add/drop multiplexers (ADMs), is a more meaningful metric for the network or topology rather than the number of wavelengths, and that the independent traffic streams that wavelength routed networks will carry are likely to have small bandwidth requirements compared even to the bandwidth available in a single wavelength of a WDM system. These two issues give rise to the concept of *traffic grooming* [6, 5, 11, 12, 2, 1, 7] which refers to techniques used to combine lower speed traffic components onto available wavelengths in order to meet network design goals such as cost minimization.

The problem of designing logical topologies for rings that minimize cost as measured by the amount of electro-optic equipment has recently received much attention in the literature [11, 12, 1, 2, 5, 6, 7]. The problem is addressed in part or full to arrive at heuristic solutions in [11, 12, 2, 7]. A common cost measure used in literature is the number of SONET ADMs

---

*This work was supported by the NSF under grant NCR-9701113.

[5, 11]. Wavelength assignment to lightpaths has been recognized to be an important part of the problem and several studies focus on this [9, 5], while others consider the lightpath routing problem as well [7]. In [11, 12], a strategy of first grooming traffic components into circles is presented; in [11] these circles are then groomed, in [12] they are scheduled in a sequence of virtual topologies. Heuristic algorithms to minimize network cost by grooming are presented in [2], for special traffic patterns such as uniform, certain cases of cross-traffic, and hub. In [1], a heuristic algorithm based on a bipartite matching formulation of the problem is presented for specific traffic characteristics.

As has been noted in literature [4, 11], the problem of logical topology design is NP-hard even for a physical ring topology, and achievability bounds are useful for evaluating performances of heuristic algorithms. We present a new framework for computing bounds for the problem of optimal traffic grooming in physical ring topologies. We decompose the ring into path segments consisting of successively larger number of nodes. We show that solving a path segment exactly is much easier than solving a ring of the same number of nodes. We combine the path solutions to obtain a series of bounds, both lower and upper. Computation of the bounds require less effort than computing the optimal value, and depending on the problem instance, several bounds in the sequence are likely to require significantly less effort.

The problem we consider is very general, as we do not impose any constraints on the traffic patterns. Furthermore, the upper bounds we derive are based on actual feasible topologies, so our algorithm for obtaining the upper bounds is a heuristic for the general problem of traffic grooming. Finally, although we illustrate our approach using a specific formulation of the problem, it is straightforward to modify it to apply to a wide range of problem variants with different objective functions and/or constraints such as multiple fiber links between nodes, physical hop limit, bidirectional rings, etc.

2 Problem Formulation

We consider a unidirectional ring $\mathcal{R}$ with N nodes numbered from 0 to $N-1$, as shown in Figure 1(a) (working ring only). The fiber link between each pair of nodes can support W wavelengths, and carries traffic in the clockwise direction; in other words, data flows from a node i to the next node $i \oplus 1$ on the ring, where $\oplus$ denotes addition modulo-N. (Similarly we use $\ominus$ to denote subtraction modulo-N.) The links of $\mathcal{R}$ are numbered from 0 to $N-1$, such that the link from node i to node $i \oplus 1$ is numbered i. Each node in the ring is equipped with a wavelength add/drop multiplexer (WADM) (see Figure 1(b)). A WADM can perform three functions. It can *add* or *drop* optical traffic from/into electronic form, for electronic processing. The processing is traffic termination or possibly electronic routing. It can also optically switch some wavelengths from the incoming link of a node directly to its outgoing link without the need for electro-optic conversion of the signal carried on the wavelength. We assume that estimates of the aggregate node-to-node traffic are available, requiring the design of a *virtual* or *logical* topology (see the discussion below) consisting of a set of static lightpaths. (We do not consider the *dynamic* case of requests for lightpaths or traffic components arriving continuously during operation, this latter problem has been considered elsewhere [13, 14, references thereof].) The traffic demands between pairs of nodes in the ring are given in the traffic matrix $T = [t^{(sd)}]$. All traffic is expressed in terms of some base rate (*e.g.* OC3) and the bandwidth available in a single wavelength is denoted by C in this unit.

Given the ring physical topology, a *logical* topology is defined by establishing *lightpaths* between pairs of nodes. A lightpath is a direct optical connection. More specifically, if a lightpath spans more than one physical link in the ring, its wavelength is optically passed through by WADMs at intermediate nodes. We assume that ring nodes are not equipped with wavelength converters, therefore a lightpath must be assigned the same wavelength on all physical links along its path.

In this paper we have chosen to concentrate on the total electronic forwarding (routing) performed by a virtual topology as the performance metric to be optimized. This is of interest because such forwarding involves electro-optic conversion and added message delay and processor load at the intermediate nodes. There is also the possibility of increased buffer requirement and queueing delay. We forbid a traffic component to be carried completely around the ring before being delivered at the destination, thus each traffic component can traverse a given link at most once. This is reasonable because the alternative consumes more bandwidth, and is likely to require more electronic routing. We can formulate the problem of designing a virtual topology for a ring network such that the total amount of electronic routing at the ring nodes is minimized as an Integer Linear Problem (ILP), following the formulation in [4] for the general topology case. The specific details and the mathematical formulation for the ring network is omitted here and can be found in [3]. It consists of $O(N^4 + N^2W)$ constraints and $O(N^4 + N^2W)$ variables, where N is the number of nodes in the ring, and W the number of wavelengths.

3 Path Decomposition of a Ring Network

Definition of Decomposition: We consider a ring $\mathcal{R}$ with N nodes labeled $0 \cdots (N-1)$, in order, and traffic matrix T. We define a *segment* of length $n, 1 \leq n \leq N$, starting at node $i, 0 \leq i < N$, as the part of the ring $\mathcal{R}$ that includes the n consecutive nodes $i, i\oplus 1, i\oplus 2, \cdots, i\oplus(n-1)$, and the links between them.

We define a *decomposition of ring $\mathcal{R}$ around a segment of length n starting at node i* as a path $\mathcal{P}_n^{(i)}$ that consists of $n+2$ nodes and $n+1$ links as follows: the n nodes and $n-1$ links of the segment of ring $\mathcal{R}$ of length n starting at node i, a new node S and a link from S to i, and a new node D and a link from node $i \oplus (n-1)$ to D. We also refer to $\mathcal{P}_n^{(i)}$ as an n-node decomposition of ring $\mathcal{R}$ starting at node i. Figure 2 shows such a decomposition.

Associated with the decomposition $\mathcal{P}_n^{(i)}$ is a new traffic matrix $T_{\mathcal{P}_n^{(i)}} = \left[t^{(sd)}_{\mathcal{P}_n^{(i)}}\right]$, $s, d \in \{i, i \oplus 1, \cdots, i \oplus (n-1), D, S\}$, derived from T, the original traffic matrix, as follows:

$$t^{(sd)}_{\mathcal{P}_n^{(i)}} = \begin{cases} t^{(sd)}, & \text{if} \quad i \preceq s \prec d \preceq i \oplus (n-1) \\ \sum_{j \notin \{i, i\oplus 1, \cdots, d \ominus 1\}} t^{(jd)}, & \text{if} \quad s = S, i \preceq d \preceq i \oplus (n-1) \\ \sum_{j \notin \{s\oplus 1, s\oplus 2, \cdots, i\oplus(n-1)\}} t^{(sj)}, & \text{if} \quad d = D, i \preceq s \preceq i \oplus (n-1) \\ t_{\text{pass-thru}}(i, n), & \text{if} \quad s = S, d = D \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where $t_{\text{pass-thru}}(i, n)$ denotes the traffic of the original matrix T that *passes through* the segment of length n starting at node i, i.e., traffic on ring $\mathcal{R}$ that uses the links of the segment but does not either originate or terminate at any of the nodes in that segment. The amount of this traffic can be readily obtained by inspection of traffic matrix T. We have used $s \prec d$ in the above expression to denote that node s precedes node d in the decomposition and $s \preceq d$ to denote that node s precedes and may be the same as node d in the decomposition.

Because of the way the traffic matrix for the decomposition is defined in (1), from the point of view of any node $k, i \preceq k \preceq i \oplus (n-1)$ in the segment, the traffic pattern in the new path $\mathcal{P}_n^{(i)}$ is *exactly the same* as in the original ring. The new nodes S and D are introduced in the decomposition to abstract the interaction of traffic components between nodes in and outside the segment. The new node S acts as the source of all traffic flowing into the segment in the real ring. Similarly, the new node D acts as the destination for all traffic seen to flow out of the segment. The details of the flow of these traffic components in the rest of the ring is hidden from the segment. Finally, the fact that $\mathcal{P}_n^{(i)}$ is a path (i.e., that there is no link from node D to node S) means that the details of traffic in the original ring that does not involve any nodes or links of the segment are hidden in the decomposition.

Solving Path Segments in Isolation: Consider the traffic matrix $T_{\mathcal{P}_n^{(i)}} = \left[t^{(sd)}_{\mathcal{P}_n^{(i)}}\right]$ of the decomposition $\mathcal{P}_n^{(i)}$ of a segment of length n starting at node i in the ring $\mathcal{R}$, as given in (1). This matrix can be thought of as representing the traffic demands in a ring network consisting of nodes $S, i, \ldots, i \oplus (n-1), D$, where there is simply no traffic flowing over the link from node D to node S. Consequently, the ILP formulation we mentioned in Section 2 can be used to obtain a virtual topology that minimizes electronic routing for this ring with traffic matrix $T_{\mathcal{P}_n^{(i)}}$. Since the ILP formulation disallows traffic routing that carries a traffic component beyond its destination and all around the ring, no lightpaths can be formed to carry traffic over the link from D to S that is absent in the decomposition. Thus, the topology obtained in this manner can be directly applied to the path $\mathcal{P}_n^{(i)}$. When we use the ILP to find an optimal topology for path $\mathcal{P}_n^{(i)}$ we will say that we solve the n-node segment *in isolation.*

The topology so obtained achieves the locally best value for electronic routing for the set of nodes in the decomposition, but does not take into account the rest of the ring. This topology may not form the subtopology for these nodes for any optimal solution, or even any feasible solution, of the complete ring. We denote the optimal objective value for the decomposition $\mathcal{P}_n^{(i)}$ by $\phi_n^{(i)}$. Since the new nodes S and D do not contribute any electronic routing, and the traffic pattern seen by the n nodes abstracted from the real ring is the same as when they are included in the ring, $\phi_n^{(i)}$ also represents the locally best (lowest) amount of electronic routing at this set of nodes when considered as part of the ring.

Our motivation for using the path decomposition described in this section is two-fold. First, as the number n of nodes in a segment starting at some node i increases, the resulting decomposition $\mathcal{P}_n^{(i)}$ will more closely approximate the original ring. As a result, the bounds we obtain will be tighter with increasing n. Second and more importantly, a path decomposition significantly alleviates the problem of exponential growth in computational requirements for solving the original ILP for an n-node network. This result is a direct consequence of the following lemma. The proof is straightforward and is based on assigning arbitrary wavelengths on the first link and extending them over the path. It is omitted here, and can be found in [3].

Lemma 3.1 *A wavelength assignment always exists for a feasible virtual topology on a unidirectional path network, and it can be obtained in time linear in the number of links and the number of wavelengths W per link.*

In solving the decomposed problem, we are merely interested in the optimum value of the objective, since this is the value from which we will obtain the bounds. Since we know that a wavelength assignment is always possible, we can eliminate the variables and constraints related to wavelength assignment from our formulation. The order of the numbers of variables and constraints remain the same, but a significant number of them are removed. This creates a formulation that is smaller and requires dramatically less computation to solve. In practice, we have found that eliminating the wavelength assignment subproblem can result in a reduction in computation time by several orders of magnitude. For instance, completely solving a six-node ring network using the original formulation (with wavelength assignment) requires between 60 and 90 minutes on a Sun Ultra-10 workstation. However, solving a six-node path network using the simplified formulation (no wavelength assignment) requires only a few seconds. In both cases, we used the LINGO scientific computation package which utilizes branch-and-bound algorithms to solve the ILP.

4 Bounds

In this section we describe how we can combine the $\phi_n^{(i)}$ values we get from n-node decompositions to obtain lower as well as upper bounds on the total amount of electronic routing performed in the optimal case by a virtual topology on the original ring. The method of combination is different for lower and upper bounds.

4.1 Lower Bound

Recall that $\phi_1^{(i)}$ represents the locally best amount of electronic routing that can be achieved at node i considered in isolation and therefore in any topology, there can be no topology which achieves an even lower value. Thus, $\phi_1^{(i)}$ is a lower bound on the amount of electronic routing performed at node i for any feasible virtual topology, and in particular, for the optimal virtual topology. Since $\phi_1^{(i)}$ represents contribution to the electronic routing only by node i, we can add the contributions together for each node to obtain a lower bound on the total electronic routing performed for all nodes in a feasible virtual topology. We denote this sum by $\Phi_1 = \sum_{i=0}^{N-1} \phi_1^{(i)}$. The quantity Φ_1 is a lower bound on the objective value (total electronic routing) for any feasible virtual topology, and in particular, for the optimal virtual topology for the ring $\mathcal{R}$.

Generalizing this notion, we find that we can add the $\phi_n^{(i)}$ values for any set of decompositions that involve segments that are disjoint in the ring, and we are still guaranteed to obtain a lower bound on the objective value for any feasible topology. We formalize this in the following lemma:

Lemma 4.1 *Let σ_n be a set of segments of ring $\mathcal{R}$ which partition the nodes of the ring in segments of length n or smaller. Let $l_k, l_k \leq n$, denote the length (number of nodes) of segment $k, k = 1, \cdots, \mid \sigma_n \mid$, and let i_k denote its starting node. The quantity*

$$\Phi(\sigma_n) = \sum_{k=1}^{|\sigma_n|} \phi_{l_k}^{(i_k)} \tag{2}$$

is a lower bound on the objective value for any feasible virtual topology on the ring $\mathcal{R}$, and therefore on the optimal objective value.

We now define Φ_n as:

$$\Phi_n = \max\{\Phi(\sigma_n)\} \tag{3}$$

where the maximum is taken over all partitions of the ring which contain segments with n or less nodes. Figure 4 shows two partitions of the same ring, the first containing only 1- and 2-node segments, and the second containing only 1-, 2- and 3-node segments. Note that our definition of Φ_1 before is consistent with this general definition.

It is obvious from definition (3) that the set of partitions we consider in computing Φ_{n+1} is a proper superset of that we consider in computing Φ_n. As a result, we have a strong sequence of bounds in which each is at least as tight as the previous one, that is, $\Phi_{n+1} \geq \Phi_n \, \forall n \in \{1 \cdots (N-1)\}$.

4.2 Upper Bound

We first note that the value of the objective function for any feasible virtual topology sets an upper bound on the optimal value, since it corresponds to an actual solution and the optimal solution can only be better than or as good as this solution. Thus, we consider different achievable topologies and we obtain upper bounds from them.

First we consider the simplest virtual topology possible, namely, the topology consisting only of W single-hop lightpaths from each node i to the next node $i \oplus 1$. All traffic not terminating at any given node must be electronically routed by that node. We will call this the *no-wavelength-routing* topology, since no wavelengths are optically routed at any node and each lightpath spans exactly one physical link. (This type of topology is called a *PPWDM ring* in [6].) In such a topology, each node i performs the maximum possible amount of electronic routing, which we denote by $\psi^{(i)}$. Quantities $\psi^{(i)}, i = 0, \cdots, (N-1)$, can be readily obtained from the traffic matrix T. We denote the total electronic routing performed under the no-wavelength-routing topology by $\Psi_0 = \sum_{i=0}^{N-1} \psi^{(i)}$. Since this is a feasible topology, Ψ_0 is an upper bound on the optimal electronic routing,

Let us call *concentrator nodes* those nodes which do not perform any optical forwarding (in the no-wavelength-routing topology, every node is a concentrator node). Nodes S and D can be viewed as concentrator nodes in the single node decomposition, thus we see that nodes of any segment preceded and succeeded by a concentrator node each is free to route traffic to achieve the local best case, corresponding to $\phi_n^{(i)}$ for that segment. For example, a topology may be created by alternating concentrator nodes with single nodes which route traffic according to the corresponding single node decomposition solutions. This topology is illustrated in Figure 3, where the even-numbered nodes are concentrator nodes. The objective value for such a topology would be an alternate sum of $\phi_1^{(i)}$ and $\psi^{(i)}$ values. For the strongest possible upper bound, we would choose the topology which minimizes this sum.

It is now straightforward to obtain a strong sequence of upper bounds along the same lines. We define Ψ_n as the lowest objective value we obtain for all topologies which are created by alternating concentrator nodes with segments no larger than n nodes in size. We can once again consider this in light of partitions of the nodes of a ring. Now, however, the partitions are constrained not only to use segments of n-nodes or less, but every alternate segment must contain exactly one node. These alternate single-node segments are used as concentrator nodes in the topology we create, rather than as single-node decompositions.

We note that since every decomposed segment has to alternate with a concentrator node, we can only use up to $N-1$ node decompositions, and cannot use N-node decompositions as for the lower bounds. As before, the set of all topologies we consider in obtaining Ψ_{n+1} is a superset of the set of all topologies we consider in obtaining Ψ_n, therefore we may assert that $\Psi_{n+1} \leq \Psi_n \, \forall n \in \{0 \cdots (N-2)\}$, giving us a strong sequence of upper bounds.

Because the bounds $\{\Psi_n\}$ are based on actual feasibly topologies, they also provide us with a useful series of heuristic solutions to the ring. Below we mention a result which shows the tightness of the bounds and thus the goodness of the heuristics, and we see in Section 5 that even the first few solutions of the series can outperform a simplistic heuristic. The later solutions in the series can compare favorably with some heuristics reported in literature. Specifically, Ψ_{N-1} must be as good or better than a *single-hub* architecture [6, 2, 1], because it considers all topologies with a single concentrator node (which is equivalent to a hub node). For a similar reason, $\Psi_k, k \geq \lceil N/2 \rceil$ must be as good or better than a double hub design, if the hubs are constrained to be diametrically opposite in the ring.

Both the upper and the lower bounds get progressively tighter with increasing n, and for the final bounds we can assert that:

$$\Psi_{N-1} - \Phi_{N-1} \leq \min_{i=0}^{N-1} (\psi^{(i)} - \phi_1^{(i)}) \tag{4}$$

The derivation is based on considering a partition of the ring into the single node which yields the minimum $\psi^{(i)} - \phi_1^{(i)}$, and the $N-1$ decomposition formed by the rest of the nodes. We then obtain both and upper and lower bound from this partition and compare them to Ψ_{N-1} and Φ_{N-1}. The derivation is omitted here and can be found in [3].

The bounds Φ_n (and Ψ_n) for successive values of n incorporate progressively more information about the problem and as such require progressively more computational effort to determine. The dominating factor in the computational effort is the calculation of the $\phi_n^{(i)}$ values required for a given bound, as we have noted before the number of variables and constraints increase as $O(n^4)$. Thus the maximum value of n for which we can determine the corresponding bounds is limited by this computational effort. The computational effort required to combine the $\phi_n^{(i)}$ values into the best available bounds is $O(2^N)$ for a straightforward algorithm, but the problem can be presented as a dynamic programming problem which takes only $O(n^2N)$ time to solve. We omit this algorithm here. The algorithm can be found in [3], as well as a lemma which can introduce a further constant order reduction in the computation required.

5 Numerical Results

We define a traffic pattern as *statistically symmetric* in which the traffic components of the form $t^{(s\oplus x, d\oplus x)}$ for all x and for any given s and d are all drawn from the same distribution. In such a case the traffic pattern looks roughly the same from all nodes of the ring. We concentrate on such traffic patterns for producing our numerical results because for highly asymmetric traffic patterns the difference in performance between the best and worst performing topologies is likely to be less. We further categorize statistically symmetric traffic matrices as rising, falling or uniform depending on whether the traffic components sourced by a node to other nodes increase, decrease or remain the same with distance to destination node. This concept is also statistical in nature. We characterize a traffic pattern by the average physical traffic load on the links of the ring (the loads on individual links are likely to be close to each other because of the symmetrical nature of the traffic) expressed as a fraction of WC, the maximum possible, for comparison purposes.

We present results pertaining to 8-node and 16-node rings. For most of our results, the value of W was taken to be between 16 and 20 and the value of C around 48. We used randomly generated statistically symmetric traffic matrices for all the runs. A discrete uniform probability distribution was used for all traffic generation. We focus on characteristic physical load values of 50% and 90%. Only a sampling of the results obtained are presented here; for more numerical results see [3].

The quantity Ψ_0 denotes the amount of electronic routing performed by a topology that does not employ optical forwarding at all. This is often actually used in networks at transitory stages [4, 8]. We can consider this case to correspond to *no grooming*, that is, no effort has been made to groom individual traffic components into lightpaths. The other extreme (not necessarily achievable) is *complete grooming*, in which all traffic is groomed into lightpaths and no electronic routing is performed. The actual amount of electronic routing performed by any feasible topology falls between these limits and may be expressed as a fraction of Ψ_0 to indicate the *effectiveness of grooming*. We express all quantities plotted below by normalizing them to Ψ_0 to express grooming effectiveness values accordingly.

Figures 5 and 6 show detailed results for 8-node rings, statistically uniform traffic with 50% load, and statistically rising traffic with 90% load respectively. Figures 7 and 8 both show detailed results for 16 node rings, statistically falling traffic with 90% load. Figure 7 represents a case where the traffic from a node falls to zero at the farthest node, whereas Figure 8 represents one in which the traffic falls to zero at a node halfway around the ring, and is zero to the farther nodes. All the results show similar characteristics. There is a sharp decrease from Ψ_0 to Ψ_1 and more moderate decrease thereafter. The quantity Ψ_7 is between 0.1 and 0.2 in all cases. We generally observe that we get good grooming effectiveness and that the lower bounds are comparatively less in magnitude. This validates our approach of describing the values of the bounds with respect to the no-optical-forwarding case rather than the optimal value, because it indicates that a high value of electronic routing for some feasible topology is more likely to result from lack of grooming (and can be corrected by proper grooming) than being the inevitable consequence of a high optimal value. We also plot the values of two lower bounds computed after the fashion of the Moore bound following [10]. These have been developed by consideration of general topologies and it is expected that our bounds, derived for the special case of the ring, will be tighter. In fact we see that, in most cases, we obtain only the trivial value of 0 for these bounds. However, for the 16-node ring in the case where traffic falls to zero at the end of the ring (figure 7), the first bound has a comparable value to the largest Φ_n we have obtained.

The other two figures represent results of 30 traffic matrices of the same traffic characteristics each. Figure 9 represents an ensemble of 30 traffic matrices of 8-node rings, each of uniform traffic with 90% load; in this case the absolute electronic routing values are plotted. Figure 10 represents an ensemble of 30 traffic matrices of 16-node rings, uniform traffic with 50% load; normalized data as with the previous results are plotted. The 2-hop lower bound is also plotted. For the latter ensemble the highest value of n for which Φ_n and Ψ_n are plotted is 5, for the earlier one it is 6. The ensemble results confirm the detailed results we obtained earlier.

We also plot an easy to compute lower bound on the performance of a simple heuristic which is based on solving the problem optimally but using only single-hop and two-hop lightpaths, a heuristic attractive for its simplicity and potential for easy OAM. A lower bound on the performance of such a heuristic is easy to obtain by considering the minimum number of times each traffic component must be electronically routed. We call this bound the *2-hop lower bound.* Since Ψ_0 and Ψ_1 are obtained from topologies that can by definition contain no lightpaths longer than two hops, the objective value of the optimal two-hop topology will by definition be equal to or less than these. However, in both detailed and ensemble results we see that most Ψ_n values for $n > 1$ are lower than the 2-hop lower bound. Thus even the first few solutions provided by our framework can outperform simplistic heuristics such as the two hop optimal topology.

6 Concluding Remarks

We have considered the problem of grooming traffic in virtual topology design for wavelength routed optical networks. We have created a framework of bounds, both upper and lower, on the optimal value of the amount of traffic electronically routed in the network. The bounds are obtained based on the idea of decomposing the ring network a few nodes at a time. We specify the decomposition method and derive a result showing that solving the decompositions is a considerably more tractable problem than solving the complete problem. We present a method of combining these partial solutions into a sequence of bounds, both upper and lower, in which every successive bound is at least as strong as the last one.

We present numerical results of the computation of these bounds by computing the bounds on different families of traffic matrices. Numerical results indicate that the expectations from theoretical considerations are fulfilled. For larger rings, the sequence cannot be computed to the end and we are limited by the availability of computing power in how far we can compute this sequence. The numerical results show that we can get good results in either case. The upper bounds are based on constructing actual feasible topologies on the network, and thus also provide us with a sequence of increasingly good heuristics. The framework can be adapted to other formulations of the problem on the ring network, and we feel this is a useful framework in a broad context of ring problems for wavelength routing optical networks.

References

[1] R. Berry and E. Modiano. Minimizing electronic multiplexing costs for dynamic traffic in unidirectional sonet ring networks. *IEEE International Conference on Communications*, 3:1724–1730, 1999.

[2] A.L. Chiu and E.H. Modiano. Reducing electronic multiplexing costs in unidirectional SONET/WDM ring networks via efficient traffic grooming. *IEEE Globecom*, 1:322–327, Nov 1998.

[3] R. Dutta and G.N. Rouskas. A sequence of bounds for the problem of minimizing electronic routing in wavelength routed optical rings. *NCSU CSC department Technical Report*, (TR-2000-11), 2000.

[4] R. Dutta and G.N. Rouskas. A survey of virtual topology design algorithms for wavelength routed optical networks. *Optical Networks Magazine*, 1(1):73–89, Jan 2000.

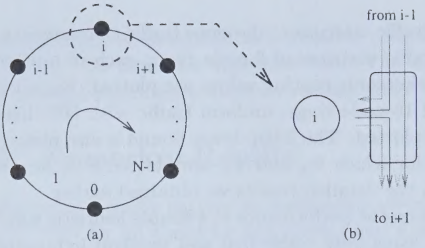


Figure 1: (a) An N-node unidirectional ring, and (b) detail of a node with a WADM

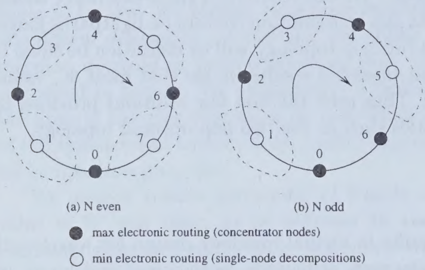

3
4
5
2
6
1
7
0
(a) N even
3
4
2
5
1
6
0
(b) N odd
● max electronic routing (concentrator nodes)
○ min electronic routing (single-node decompositions)

Figure 3: Virtual topology with alternating single-node decompositions and concentrator nodes

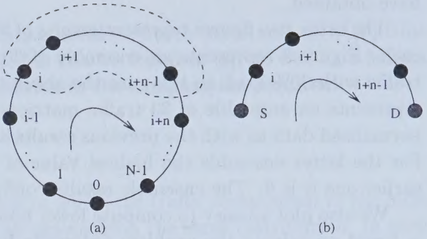


Figure 2: An n-node decomposition: (a) original ring $\mathcal{R}$, and (b) a decomposition $\mathcal{P}_n^{(i)}$

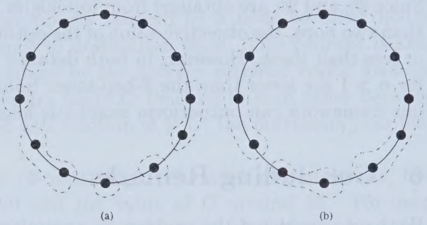


Figure 4: Partitions of the nodes of a ring into (a) segments of no more than 2 nodes, and (b) segments of no more than 3 nodes

[5] O. Gerstel, P. Lin, and G. Sasaki. Wavelength assignment in a WDM ring to minimize cost of embedded sonet rings. *Proc. IEEE INFOCOM*, pages 94–101, 1998.

[6] O. Gerstel, R. Ramaswami, and G. Sasaki. Cost effective traffic grooming in WDM rings. *Proc. IEEE INFOCOM*, pages 69–77, 1998.

[7] G. Li and R. Simha. Efficient routing to reduce the cost of add-drop multiplexers in WDM optical ring networks. *Proceedings of the SPIE*, 3843:258–267, 1999.

[8] B. Mukherjee, S. Ramamurthy, D. Banerjee, and A. Mukherjee. Some principles for designing a wide-area optical network. *IEEE/ACM Trans. Networking*, 4(5):684–696, Oct 1996.

[9] R. Ramaswami and G.H. Sasaki. Multiwavelength optical networks with limited wavelength conversion. *Proc. IEEE INFOCOM*, pages 489–498, 1997.

[10] R. Ramaswami and K.N. Sivarajan. Design of logical topologies for wavelength-routed optical networks. *IEEE JSAC*, 14(5):840–851, June 1996.

[11] X. Zhang and C. Qiao. An effective and comprehensive solution to traffic grooming and wavelength assignment in SONET/WDM rings. *SPIE Conference on All-Optical Networking:Architecture, Control and management Issues*, 3531:221–232, Nov 1998.

[12] X. Zhang and C. Qiao. On scheduling all-to-all personalized connections and cost-effective designs in WDM rings. *IEEE/ACM Transactions on Networking*, 7(3):435–445, Jun 1999.

[13] Y. Zhu, G.N. Rouskas, and H.G. Perros. Blocking in wavelength routing networks, part I: The single path case. *Proceedings of IEEE Infocom '99,New York, NY*, pages 321–328, March 1999.

[14] Y. Zhu, G.N. Rouskas, and H.G. Perros. Blocking in wavelength routing networks, part II: Mesh topologies. *Proceedings of the Sixteenth International Teletraffic Congress (ITC 16), Edinburgh, UK*, pages 1321–1330, Jun 1999.

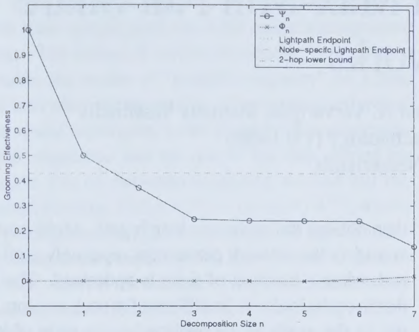


Figure 5: Detailed results for $N = 8$, Statistically uniform pattern, 50% load

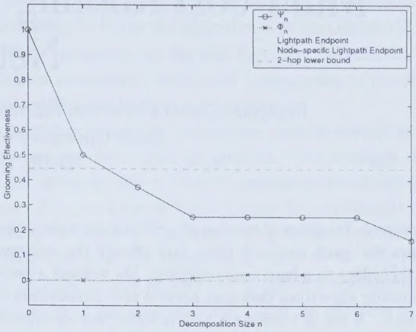


Figure 6: Detailed results for $N = 8$, Statistically rising pattern,90% load

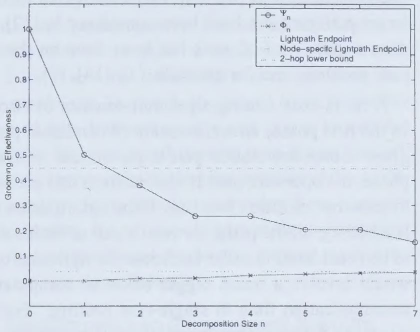


Figure 7: Detailed results for $N = 16$, Statistically falling pattern, 90% load, falling to end

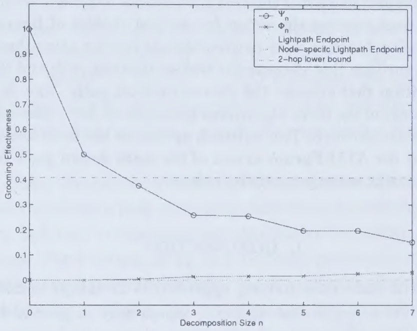


Figure 8: Detailed results for $N = 16$, Statistically falling pattern, 90% load, falling to $N/2$

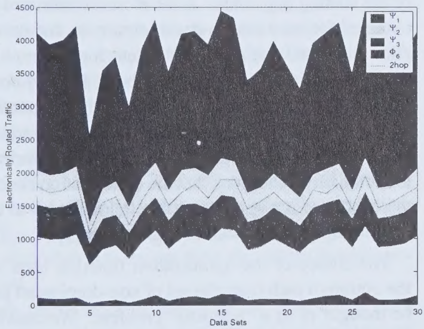


Figure 9: Ensemble results for $N = 8$, Statistically uniform pattern, 90% load, electronic routing

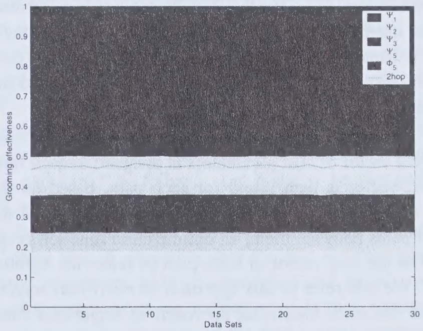


Figure 10: Ensemble results for $N = 16$, Statistically uniform pattern, 50% load, (normalized)

Multi-cost Routing in Max-Min Fair Share Networks

Francisco Javier Gutierrez, Emmanouel A. Varvarigos, Stamatis Vassiliadis
Delft University of Technology (TU Delft)
Delft, The Netherlands

Abstract—**Traditional routing algorithms use a single cost measure for each network link, and choose the optimum path according to a least cost criterion. We present a multi-cost routing algorithm that uses several cost parameters for each link. A set of candidate paths, called non-dominated paths, is obtained, and the optimum path is chosen according to an optimization function. The only limitation on the optimization function is that it is assumed to be monotonic with respect to each of the cost parameters. We compare the multi-cost routing algorithm for several choices of the optimization function with two traditional routing algorithms: the algorithm that chooses the widest-shortest path and the algorithm that chooses the shortest-widest path. The performance of the three algorithms is compared for a max-min fair share network. This network operation has been identified by the ATM Forum as one of the main design goals for ABR traffic management algorithms.**

I. INTRODUCTION

THE traditional routing approach is to assign to each link a single cost measure, which may in general be a function of several network parameters, and execute an algorithm to obtain the minimum cost path [15]. Routing algorithms of this kind limit the type of cost criteria that can be used for selecting a path, and do not permit different sessions to be treated in different ways, depending on the quality of service they require. Moreover, a single path is usually obtained for each source-destination pair, which leads to non-uniform distribution of traffic and the potential of oscillation problems [2].

In this paper we examine the more general approach, where a vector of cost parameters is assigned to each link. A cost vector is then found for each path, based on the cost vectors of its links. Given a pair of nodes and the set of paths between them, an optimization function is applied to the cost vector of each path to select the optimal path. We will refer to this approach as *multi-cost routing* and to the more traditional approach as *single-cost routing*. Although the two approaches may look similar, they are actually rather different. Single-cost routing combines the different parameters for a link in a single cost parameter for that link, often called the length of the link, and then obtains the minimum length path. Multi-cost routing considers the network parameters separately until the very end, when a function of them is optimized. The function that is optimized can be different for each session, depending on the needs of the application. In spite of its potential advantages, multi-cost routing has been rarely studied. General multi-cost routing with two cost parameters has been considered in [8], [9]. Particular cases of the bicriterion path problem have been considered in [12], [1], [4]. Also a great deal of work has been done on the quickest path problem; see, for example, [3], [14], [7].

A multi-cost routing algorithm consists of two phases. In the first phase, an enumeration of candidate paths for a given source-destination pair is performed. In the second phase, the optimum path is chosen from this set according to some optimization function. If the optimization function is arbitrary, all the paths between a pair of nodes may need to be considered in order to choose the optimum one. This would involve a much bigger effort in computation and communication than in single-cost routing. Fortunately, when the function that is optimized is monotonic in each of its parameters, as is the case with most cost functions of interest, the number of paths that are considered is greatly reduced by discarding some of them at intermediate phases of the routing algorithm. A set of non-dominated paths is obtained for each source-destination pair, and an optimization function (that may be different for different connections) is applied to select one of them. By *non-dominated path*, we mean that it is not possible to find another path that is better with respect to one of the cost criteria without being worse with respect to some other criterion. The monotonicity conditions that the optimization function has to satisfy are not very restrictive, allowing us to consider the routing problem with enough generality.

The choice of the optimization function used to select the optimum path from the set of non-dominated paths can be thought of as a "political" problem. We consider several choices of the optimization function and compare their performance for max-min fair share networks. In max-min fair share networks all the connections using a link obtain a "fair share" of the link capacity. The concept of max-min fair share has been identified by the ATM forum as one of the main design goals for ABR traffic management algorithms. The nature of bandwidth sharing in such networks makes the notion of "available capacity" on a link different than what we are used to. We examine both regular and irregular topologies, under a variety of assumptions on the link capacities and the size of the data carried by each session. For the networks examined, we find that the optimization function $f(d,b,r,h) = (d+b/r) \cdot h^{0.5}$, where d is the propagation delay, r is the available (in the max-min fair share sense) capacity, b is the amount of data to be transfered and h is the number of hops, outperforms the widest-shortest and the shortest-widest routing algorithms as well as other multi-cost routing algorithms, by taking into account both the interests of the network and the interests of the user. We also investigate the effect that long propagation delays and outdated link utilization information have on the performance of the routing algorithms.

The remainder of the paper is organized as follows. In Section 2 we give some definitions and obtain some preliminary results. In Section 3 we present some properties of the set of non-dominated paths and give a formal description of a Dijkstra-based algorithm to obtain this set. In Section 4, we consider the implementation of the multi-cost routing algorithm. Finally, in Section 5 we compare multi-cost routing for several choices of the optimization function to two traditional single-cost routing algorithms for max-min fair share networks.

II. DEFINITIONS AND PRELIMINARY RESULTS

A *network* is represented by a *directed graph* $G = (N, A)$, where N is a set of nodes and A is a set of ordered pairs of nodes. A *path* in the network is defined as a sequence of distinct nodes $(n_1, n_2, \ldots, n_l)$. Each link j of the network is assigned a k-dimensional cost vector $\mathbf{v_j} = (v_{1j}, v_{2j}, \ldots, v_{kj})$, where $v_{ij} \geq 0$ for all i. Each path is assigned a cost vector $\mathbf{V} = (V_1, V_2, \ldots, V_k)$ whose components are obtained from the link cost components according to some *monotonic associative operator* $\oplus$, that is, $V_i = \oplus_{j=1}^{l} v_{ij}$. By monotonic we mean that either $a \oplus b \geq a$ for all $a, b \geq 0$, or $a \oplus b \leq a$ for all $a, b \geq 0$ (but not both). The *cost of a path* is given by a function $f(V_1, V_2, \ldots, V_k)$ of its cost vector. The objective of the routing algorithm is to choose the path that has the the minimum cost.

We are particularly interested in the case where each cost component V_i of a path made up of l links is obtained from the link cost components in one of the following two ways:

- $V_i = \sum_{j=1}^{l} v_{ij}$, $v_{ij} \geq 0$, and f is monotonically increasing in V_i (so our objective is to minimize V_i), or
- $V_i = \min_{j=1,\ldots,l}\{v_{ij}\}$, $v_{ij} \geq 0$, and f is monotonically decreasing in V_i (so our objective is to maximize V_i).

Cost parameters of the first type will be referred to as *additive cumulative*, while cost components of the second type will be referred to as *restrictive*.

Additive cumulative parameters include several important cost measures used in practice. For example, if v_{ij} is the delay on link j, then V_i represents the delay of the path. If $v_{ij} = 1$ for all links j, then V_i corresponds to the number of hops on the path. If $v_{ij} \in [0, 1]$ is the probability that link j is operational, then $V_i = \prod_{j=1}^{l} v_{ij}$ is the probability that all links on a path are operational (assuming links fail independently), and the problem of finding the most reliable path can be reduced to a problem involving cumulative additive components by defining new cost components $v'_{i1} = -\ln v_{i1}, \ldots, v'_{il} = -\ln v_{il}$, where $v'_{ij} \geq 0$. Restrictive cost parameters appear in routing problems when capacities or transmission rates are considered. In particular, if v_{ij} is the available capacity on link j, then $V_i = \min_{j=1,\ldots,l}\{v_{ij}\}$ represents the capacity of a path. For the routing algorithm to favor less congested paths, the cost function f should be a decreasing function of V_i.

Other choices for the monotonic associative operator $\oplus$ used to obtain a path cost component from the corresponding link cost components are also useful in routing problems. For example, if v_{ij} is a Boolean parameter representing some property of link j (such as, that the link is operational, or reliable, or congested, or that its end-nodes can handle a certain protocol), then the OR, AND, and other Boolean operators (but not the exclusive OR operator, which is not monotonic) could be used to calculate an appropriate cost parameter for the path. We will not discuss explicitly Boolean operators, but the algorithms that we will give can be applied to such cases as well.

We denote by P the *set of all the possible paths* between a given pair of nodes. For a given cost function $f : R^k \rightarrow R$ we define the *set $P_{opt} \subset P$ of optimum paths* between the nodes, as the set of all paths that minimize f. It is also useful to define a *domination relation* among the paths in P as follows. Let p and p' be two paths in P, and $\mathbf{V} = (V_1, V_2, \ldots, V_k)$ and $\mathbf{V}' = (V'_1, V'_2, \ldots, V'_k)$ be their associated cost vectors, respectively. We assume, without loss of generality, that the first s components of these vectors are additive cumulative and the $k - s$ following components are restrictive; f is then assumed to be an increasing function of the first s components and a decreasing function of the last $k - s$ components. We say that path p dominates path p' if $V_1 \leq V'_1, \ldots, V_s \leq V'_s$, and $V_{s+1} \geq V'_{s+1}, \ldots, V_k \geq V'_k$, and the inequality holds at least once. It is clear that if p dominates p', the cost of p

will be smaller than that of p'.

We define the *set Q of non-dominated paths* between two nodes as the subset of P that is obtained once all the dominated paths have been eliminated. In other words, it is not possible to find in Q a pair of paths that verify the domination relation previously described.

Lemma 1: The set P_{opt} of optimum paths between two nodes is contained in the set Q of non-dominated paths, that is, $P_{opt} \subset Q$.

Since all the optimum paths are included in the set Q of non-dominated paths, the multi-cost routing algorithm can restrict the search for optimum paths to this set.

III. OBTAINING THE SET OF NON-DOMINATED PATHS

In this section we present an algorithm for obtaining the set of non-dominated paths from a given source node to all destination nodes. Before doing so, it is useful to describe some properties of the set Q of non-dominated paths.

A. *Some Properties of the Set of Non-Dominated Paths*

We let P be the set of all the paths between two particular nodes, and $Q \subset P$ be the set of all the non-dominated paths in P. We consider the case where the cost vector $\mathbf{V} = (d, c)$ consists of only $k = 2$ components, where d is additive cumulative, while c is restrictive. We represent the path cost vector $\mathbf{V} = (d, c)$ as a point in the 2-dimensional space, and the set P as a set of such points, as shown in Fig. 1a. Recall that a path p with cost vector $\mathbf{V} = (d, c)$ dominates a path p' with cost vector $\mathbf{V}' = (d', c')$ if $d \leq d'$ and $c \geq c'$ and the inequality holds at least once. Therefore, if we imagine the point representing a path as the center of a four quadrants graph, this path dominates all the paths in the lower right quadrant, which can then be eliminated. The set Q of non-dominated paths can be obtained graphically by successively using this rule, as shown in Fig. 1b.

Lemma 2: Let P_{dmin} be the set of paths in P that minimize the first component d, and $p_{dmin} \in P_{dmin}$ be the path that maximizes the second component c of the cost vector. Let also P_{cmax} be the set of paths in P that maximize the second component c, and $p_{cmax} \in P_{cmax}$ be the path that minimizes the first component. Then, $p_{dmin} \in Q$ and $p_{cmax} \in Q$.

Figure 1 shows graphically that both p_{dmin} and p_{cmax} belong to Q. The following lemma describes the transitive property of the domination relation.

Lemma 3: If path p dominates p' and p' dominates p'', then p dominates p''.

Note that the path in Q that minimizes the additive component is also the one that minimizes the restrictive component, and the path that maximizes the restrictive component is also the path that maximizes the additive component. Also, the paths in Q form an increasing succession in both components. These properties can be generalized to the case where $k > 2$, and will be the basis for the algorithm given in the next section for obtaining the set Q of non-dominated paths.

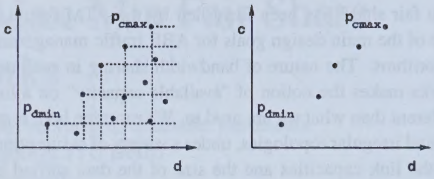


Fig. 1. Graphic representation of the paths: (a) The set P of all the paths and the way the set Q is obtained. (b) The set Q of non-dominated paths. Paths p_{dmin} and p_{cmax} are contained in it.

B. *Algorithm for Computing the Set of Non-dominated paths*

In this section we present an algorithm for obtaining the set of non-dominated paths from a given node to all other nodes of the network. The algorithm can be viewed as a generalization of Dijkstra's algorithm [5], which only considers networks with single costs. Dijkstra's algorithm requires all arc lengths to be nonnegative. Our algorithm inherits this requirement, which is consistent with the assumptions made in Section 2 for the cost vector components.

To clarify the operation of the algorithm we first describe the case where the cost vectors consist of only $k = 2$ components: the delay of the link (which is additive cumulative) and the capacity of the link (which is restrictive). The algorithm starts by considering the neighbors of the origin node, and choosing the one with the smallest link delay; in case of a tie, the one with the largest capacity is chosen. We have already obtained the first non-dominated path between the origin and that neighbor. Clearly, this path is non-dominated because the delay is an additive cumulative parameter and no other path can have a smaller delay. The algorithm then considers the neighbors of these two nodes and chooses the one that has the smallest delay to the origin and so on.

Before giving a formal description of the algorithm, we present the operation of the algorithm for the graph of Fig. 2. The evolution of the algorithm for origin node A is shown in the table of Fig. 3. Each row of the table corresponds to a destination node and the columns represent the steps of the algorithm. The first column shows the initial situation. Odd columns in Fig. 3 contain the temporary sets of non-dominated paths. Only the underlined paths are guaranteed to be non-dominated at the end of the algorithm. The other paths may be dominated and discarded in the following steps of the algorithm. The sign "+" used in the calculations does not represent addition, but the way the cost of the concatenation of a path and a link is obtained, with each component of the cost vector calculated according to the corresponding monotonic associative operator. At each step, the new cost vectors are compared to the previously obtained vectors for that node using the domination relation and any dominated paths are discarded. The last column shows the final set of non dominated paths for every destination node. The same principle and similar data structures can be used to obtain algorithms for the case where k-dimensional cost vectors are considered.

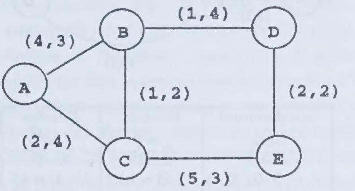


Fig. 2. Each edge is characterized by a 2-dimensional cost vector, the first component of which is additive cumulative, while the second is restrictive. An edge represents a pair of arcs with opposite directions and the same cost vector.

B	(4,3,A)	(1,2)+(2,4) (3,2,C)	(3,2,C) (4,3,A)		(3,2,C) (4,3,A)		(3,2,C) (4,3,A)	
C	(2,4,A)		(2,4,A)		(2,4,A)	(1,2)+(4,3) (5,2,B)	(2,4,A)	
D	(∞,0,-)		(∞,0,-)	(1,4)+(3,2) (4,2,B)	(4,2,B)	(1,4)+(4,3) (5,3,B)	(4,2,B) (5,3,B)	
E	(∞,0,-)	(5,3)+(2,4) (7,3,C)	(7,3,C)		(7,3,C)		(7,3,C)	(2,2)+(4,2) (6,2,D)
B	(3,2,C) (4,3,A)		(3,2,C) (4,3,A)		(3,2,C) (4,3,A)		(3,2,C) (4,3,A)	
C	(2,4,A)		(2,4,A)	(5,3)+(6,2) (11,2,E)	(2,4,A)		(2,4,A)	
D	(4,2,B) (5,3,B)		(4,2,B) (5,3,B)		(4,2,B) (5,3,B)	(2,2)+(7,3) (9,2,E)	(4,2,B) (5,3,B)	
E	(6,2,D) (7,3,C)	(2,2)+(5,3) (7,3,B)	(6,2,D) (7,3,C)		(6,2,D) (7,3,C)		(6,2,D) (7,3,C)	

Fig. 3. Evolution of the algorithm to obtain the set of non-dominated paths.

The algorithm to obtain the set of non-dominated paths between the origin and all other nodes of the network is now formally described. We assume that the link cost vectors have k components, the first s of which are additive cumulative and have to be minimized, and the remaining $k-s$ are restrictive and have to be maximized. We denote by e_{ij} the cost vector of link (i,j). Every path is represented by a label that includes the cost vector associated with it and the first hop to the origin using that path. The origin is taken to be node 1.

We let W_i be the set of labels of destination node i, and $W = \cup_{i\neq 1} W_i$ be the set of all labels. Initially, every node has a single label corresponding to the link that connects it directly to the origin node. If a node has no link connecting it to the origin, its cost vector has the first s components equal to infinity and the remaining $k-s$ components equal to zero. In each the following steps, the algorithm marks labels (equivalently, paths) from the the set W as *final*. We let $W^d \subset W$ be the subset of all final labels for all the destinations, and $W_i^d \subset W_i$ be the set of final labels for destination node i. We also let $T \subset N - \{1\}$ be the set of destination nodes with at least one final label. The algorithm can now be described as follows:

- Step 0 (**Initialization**): $W = \{2,3,\dots,N\}$, $W^d = \emptyset$, $T = \emptyset$.
- Step 1 (**Choosing the optimum label**): The label whose cost vector minimizes the first component is chosen. In case of a tie, the vector with the minimum second component is chosen. In case there is still a tie, the third components are compared and so on. For the first s components, the smallest values are chosen; for the remaining $k-s$ components, the largest values are preferred. If $\mathbf{V}$ is the cost vector of the chosen label and i is the corresponding node, then the following updates are performed:

$$W_i^d := W_i^d \cup \{\mathbf{V}\}, \quad W^d := W^d \cup \{\mathbf{V}\}, \quad T := T \cup \{i\}.$$

- Step 2 (**Obtaining the new labels**): The neighbors of node i, which may or may not belong to the set T, are now considered and are given new labels (except for the origin node and the node specified as the previous node in the label). The new label for a neighbor j of node i is then computed in the following way. The new cost vector is updated according to $\mathbf{U} = \mathbf{V} + e_{ij}$, where "+" represents the operation required to compute the new cost vector, which is componentwise addition for for the first s (additive) components, and componentwise minimization for the remaining $k-s$ (restrictive) components. The previous node on the path corresponding to vector $\mathbf{U}$ is set to be node i.
- Step 3 (**Discarding dominated paths**). Every neighbor considered in step 3 compares its new label with its previous labels using the domination relation. Let j be one of the neighbors of node i, $\mathbf{U}$ the new cost vector and W_j be the set of labels for this node. The new label has to be compared with the labels in W_j (both final and non-final). If some cost vector in W_j dominates $\mathbf{U}$, then $\mathbf{U}$ is discarded and W_j does not change; otherwise, $\mathbf{U}$ is added to the set W_j and W, so that $W_j := W_j \cup \{\mathbf{U}\}$ and $W := W \cup \{\mathbf{U}\}$. If the new vector dominates one of the vectors in W_j, then the dominated vectors are eliminated from W_j and W. Note that it is not possible for the new vector to dominate an existing vector and be dominated by

another one at the same time.

- Step 4. (**Termination**). If $W^d = W$, the algorithm is completed. Otherwise, (that is, when there are still some labels to be chosen) the algorithm goes back to Step 1.

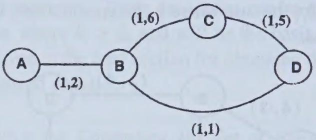
IV. IMPLEMENTATION OF THE MULTI-COST ROUTING ALGORITHM

We distinguish between two kinds of networks: networks that use distributed routing and networks that use source routing. In distributed routing every switch in the path participates in the routing decision and the switches use the packet destination and a local routing table to decide the next hop to the destination. In source routing, each packet carries a description of the entire sequence of output ports that have to be traversed.

Distributed routing algorithms are usually based on the shortest path idea, where a single cost is assigned to each link. One of the reasons that make the combination of distributed routing and shortest path routing possible is that the shortest paths satisfy a *property of inclusion*: every subpath of a shortest path is a shortest path. In other words, if a shortest path from node A to node E traverses nodes A, B, C, D, E, then a shortest path from node B to node E traverses nodes B, C, D, E. As a result, a packet sent from node A to node E only has to carry information about the destination node, and can use the forwarding tables at intermediate nodes (which are consistent with each other because of the inclusion property, assuming no changes in the link costs occur while the packet is in transit) to find the next hop.

Multi-cost routing is not always compatible with distributed routing, because for some choices of the cost function the optimal paths do not have the inclusion property that shortest paths have. To see this, consider the example of Fig. 4, where the path cost vector is $\mathbf{V} = (h, c)$. The first link cost component is equal to 1, is additive cumulative, and will account for the number of links on a path. The second component represents the link capacity, and is restrictive. We assume that each node has already computed the set of non-dominated paths for each of the other nodes, and node A is about to establish a connection to destination node D. The table of Fig. 4 shows the set of non-dominated paths from node A to node D and from node B to node D.

Consider first the optimization function $f_1 = h/c$, where h is the total number of hops and c is the minimum of the capacities on the path. This function favors paths that have a small number of links and large capacity. For this function, the optimum path from node A to node D is (A, B, C, D), while the optimum path from node A to node B is (B, C, D). Therefore, in this case the optimum path from node B to node D is contained in the optimum path from node A to node D, and the packets sent by node A to node D through node B only have to have to carry with them information about the destination node D, and not a full description of the path. Consider now the optimization function $f_2 = h^2/c$ which penalizes long paths more severely. For this function, the optimum path from node A to node D is (A, B, D) while the optimum path from node B to node D is (B, C, D). Node A can send packets to node D through node B expecting them to be routed directly to node D, but node B will route these packets through node C because that is its optimum path to node D.



Non dominated paths	Function $f = \frac{h}{c}$	Function $f = \frac{h^2}{c}$
A→D	A→D	A→D
$(A, B, C, D), (3, 2)$	$f(3, 2) = 1.5$ *	$f(3, 2) = 4.5$
$(A, B, D), (2, 1)$	$f(2, 1) = 2$	$f(2, 1) = 4$*
B→D	B→D	B→D
$(B, C, D), (2, 5)$	$f(2, 5) = 0.4$*	$f(2, 5) = 0.8$*
$(B, D), (1, 1)$	$f(1, 1) = 1$	$f(1, 1) = 1$

Fig. 4. The table of non-dominated paths for source nodes A and B and destination node D for two different cost functions. Optimum paths are marked with the symbol "*".

The previous example shows that a distributed implementation of a multi-cost routing algorithm is not always possible. Depending on the optimization function used, some intermediate nodes may not choose the paths originally intended by the source node, leading to a certain loss of optimality. This problem does not arise when source routing is used. With source routing, however, we lose the benefits that may be obtained through the more recent link utilization and topology information available at intermediate nodes.

V. PERFORMANCE RESULTS

In this section we use simulations to compare the performance of several multi-cost routing algorithms with that of two traditional single cost routing algorithms: (a) The *widest-shortest path* routing algorithm, where the path with the minimum number of links is chosen; in case of a tie, the one with the maximum capacity is selected. (b) The *shortest-widest path*, where the path with the maximum available capacity is chosen; in case of a tie, the path with the minimum number of links is selected.

The routing algorithms will be compared for networks of varying topology and link capacities, under the assumption that sessions share the links according to a max-min fair share scheme. The concept of max-min fair share was first proposed by Jaffe [10] to address the issue of fair allocation of resources in networks using virtual circuit switching, and has received considerable attention recently, since it was identified by the ATM Forum as one of the main design goals for ABR traffic management algorithms.

In a max-min fair share network, the connections using a particular link get a "fair share" of the link bandwidth. If a connection cannot use up its fair share bandwidth, because it has a lower source rate or it is assigned a lower bandwidth on another link, the excess bandwidth is split "fairly" among all other connections (see [2], [13]). In the basic definition of `fair share", each of the N connections competing for the link or excess bandwidth gets $1/N$ of the bandwidth. Other definitions such as weighted fair sharing and multi-class fair sharing, require pricing or administrative policies to assign a weight or class for each connection. Since these policies are still an active area of research, the basic max-min fair share model will be the first one to be supported by commercial ATM networks, and it is the one we use in this paper.

We assume that connection requests (sessions) are generated at each node according to a Poisson random process of rate λ, and session destinations are uniformly distributed over all nodes, except for the origin. When a session is generated it also specifies the amount of data to be transmitted. A new session requests from the network a PCR (Peak Cell Rate) and a MCR (Minimum Cell Rate). The user agrees not to transmit at rates above the PCR, and the network agrees to always provide at least the minimum cell rate MCR, which is typically the minimum required to keep the application running. An application is not required to specify peak and minimum cell rates, but it may simply accept the network default settings. In this case, the PCR defaults to the access rate and the MCR defaults to zero. A session is refused if the network cannot provide the requested MCR. In that case, a new Poisson process is generated for the session retries at a node, whose mean interarrival times are several times larger than the mean interarrival times of the original Poisson process. A session that has been blocked retries until the connection is accepted. We assume well behaved traffic sources: When the network report changes in link utilization, the sources adapt their transmission rates exactly and instantaneously. In the intervals between such changes, we assume that the sources transmit packets at a constant rate. These assumptions guarantee that the network switches do not reveive more data than they can send to their outgoing links, so that little storage is needed and the packet queueing delays at intermediate nodes can be ignored. The assumption of instantaneous feedback is clearly unrealistic, but we believe that it does not affect the comparisons between the routing algorithms. The effects of delayed feedback will be examined in Section 5.5.

The path cost parameters we consider in our simulations are the delay, the number of links, and the available capacity of the path. The delay of the path is equal to the sum of the link propagation delays, since the queueing delays in our model are negligible (the propagation delay per km of the fiber is 5μsec/km.) In reservation-based networks, the capacity of a link is usually taken to be the residual link bandwidth (unreserved bandwidth), and the capacity of a path is the minimum of the capacities of the links in the path. In the max-min fair share networks that we consider, however, the residual bandwidth cannot be used as a measure of the available capacity since the nature of bandwidth sharing in this kind of networks is different. Instead, an estimate of the max-min fair rate that would be obtained by the new connection if admitted is used as the available capacity, as will be explained in detail later.

We examine several choices for the optimization function in our simulations. The first function is $f(h,r) = \frac{h}{r}$, where h is the number of links and r is the available capacity of the path. This function tries to minimize h and maximize r, and represents a trade-off between the interests of the network and the interests of the connection. An individual connection wants to get the maximum capacity possible, independently of the number of links. The network has to prevent the uncontrolled use of the resources, so that it is fair to connections that may arrive later. This objective can be achieved by making the cost function an increasing function of the number of links. Another function we consider is $f(d,b,r) = d + \frac{b}{r}$, where d is the path propagation delay, b is the amount of data to be transmitted and r is, again, the available capacity. This function tries to minimize the total connection time, which consists of the propagation delay and the transmission time. It chooses the quickest path to transmit the data. A combination of the two previous functions is $f(d,b,r,h) = \left(d + \frac{b}{r}\right) * h$, which can be thought of as a way to obtain a quick path while penalizing paths that have an excessive number of links. It is a trade-off between the time requirements of the connection and the network general interests. The relative importance of these two considerations can be adjusted by using a function $f(d,b,r,h,n) = \left(d + \frac{b}{r}\right) * h^n$, for some parameter n.

We have developed an event-driven simulator that allows us to choose the routing algorithm, select the network topology and the link capacities, adjust the offered load through the session arrival rate and the amount of data in each session, and determine the type of updating for the routing information (instantaneous or periodic updating).

In the simulations, unless stated otherwise, the value for the available capacity is the estimated rate for the new connection obtained by running the max-min fair share algorithm of [13]. The value of PCR is assumed to be always equal to the maximum link capacity in the network. The link lengths are uniformly distributed in the interval between 50Km and 100Km, corresponding to propagation delays in the interval (0.25, 0.5)msec. Sessions in our simulations are divided in two categories depending on the amount of data to be transmitted: a) medium-size sessions, where the amount of data sent is uniformly distributed in the interval (50Kytes, 5Mbytes), and b) large-size sessions, where the amount of data sent is uniformly distributed in the interval (5Mbytes, 500Mbytes). Connections transmit data in the intervals (50Kbytes, 5Mbytes) and (5Mbytes, 500Mbytes), with 50% of the data belonging to each traffic category.

A. Simulation Results

We first consider the case where the sessions do not specify a value for MCR, in which case MCR defaults to zero, and all sessions are accepted. The efficiency of the routing algorithms can then be evaluated based on the average connection time, which is the sum of the propagation delay d and the transmission time b/r.

We consider two different topologies: (a) a 4x4 mesh network (see Fig. 5a), with $N = 16$ nodes and $L = 24$ links, and (b) a random network with $N = 10$ nodes and 3 links per node (see Fig. 5b). All links in the random network have capacity equal to 155Mbps, while links in the mesh network have capacity equal to 155Mbps or 622Mbps. Our intention in considering the random network is to see how the irregularities in the topology affect the performance of the routing algorithms. We consider 3 different mesh networks that differ on the total number of links with capacity 622Mbps: 1) a mesh with only 2 links of capacity 622Mbps, 2) a mesh with 6 links of capacity 622Mbps, 3) a mesh with 10 links of capacity 622Mbps. Our intention in considering these three cases is to see how the behavior of the routing algorithms is affected by the lack of regularity in the capacity assignment. For the topologies considered and the values assumed for the propagation delays, link capacities and amounts of transmitted data, transmission times dominate the total connection times. In Section 5.6 we discuss the case of very high speed networks, where the propagation delays are not negligible.

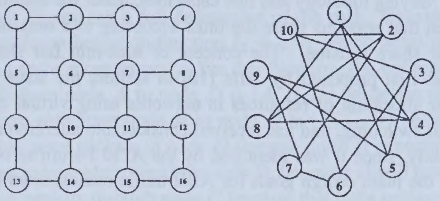


Fig. 5. (a) A 4x4 mesh. (b) The random network.

Figure 6 illustrates the connection time versus the offered load for the Mesh No. 1 network, which is very regular both in the distribution of the links and also in the values of link capacities. For light load all three algorithms have similar behavior. As the load increases, the multi-cost routing algorithm slightly outperforms the widest-shortest algorithm, because it also takes into account the parameter r, which captures the differences in the distribution of the traffic. The shortest-widest algorithm presents the worst results for heavy load. This is because it always selects the path with the maximum capacity, independently of the number of links, which leads to inefficient use of the resources. By permiting new connections to be greedy in the amount of resources they use, this algorithm reduces the amount of resources available for future connections that will consequently have larger connection times. Moreover, the longer the path is, the larger is the probability that the connection's share of the capacity will be reduced as a result of fair sharing with connections that are added later.

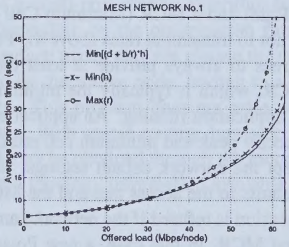


Fig. 6. Average connection time in the mesh network No. 1.

Figure 7 shows the connection time as a function of the offered load for Mesh No. 2. This mesh has a larger number of 622Mbps links than Mesh No. 1, and the differences in link capacities are more pronounced than in Mesh No. 1. For light load, all three algorithms give similar results, with the shortest-widest algorithm giving a slightly smaller average connection time. This is because in trying to maximize exclusively the capacity a session can take advantage of the high capacity of some of the links, improving the connection times. In other words, under light load conditions a session can be greedy without punishing other connections. As the offered load increases, the differences become smaller, and at heavy load the shortest-widest algorithm becomes again the worst performer.

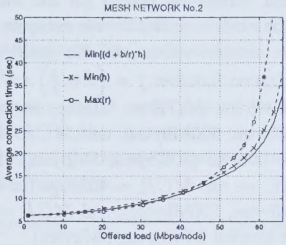


Fig. 7. Average connection time in the mesh network No. 2.

Figure 8 illustrates the connection time versus the offered load for Mesh No. 3, where the percentage of 622Mbps links is even larger than in Mesh No. 2. The shortest-widest algorithm gives the best results under light and medium load conditions. The multi-cost routing algorithm gives the second best results, and the widest-shortest algorithm gives the worst results, because it is almost insensitive to the remarkable differences in the link capacities. For heavy load, the shortest-widest algorithm gives again the worst results (see Fig. 8b).

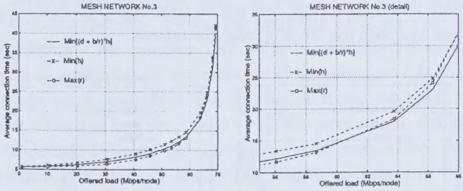


Fig. 8. (a) Average connection time for the mesh network No. 3. (b) Detail of the curves at heavy load.

These first results for the mesh network bring some interesting conclusions. The widest-shortest algorithm captures the regularity of the network and brings out acceptable results. It is essentially a non-adaptive routing algorithm that takes always the same decisions independently of the offered load. The multi-cost routing algorithm obtains always better results than the widest-shortest algorithm. This is because it captures not only the regularity of the network but also its dynamics as given by the available capacity. The efficiency of the shortest-widest algorithm depends strongly on the offered load. For light load conditions it performs well, but for heavy load it always gives the worst results. This was expected, since when the network is congested, using as little resources as possible is the main consideration.

Figure 9 illustrates the connection time versus the offered load for the random network. The widest-shortest algorithm has the worst performance. The lack of strict regularity in the network penalizes this algorithm, because it does not take into account the irregular distribution of traffic. Despite its randomness, however, the network still has a certain degree of regularity (it guarantees 3 links per node), and this is why the performance of the widest-shortest algorithm is not even worse. The shortest-widest algorithm performs slightly better for light load, but its performance gets worse as the network load increases. The multi-cost routing algorithm has good performance for all network loads, because it captures both the slight regularity in the network and the differences in the distribution of traffic.

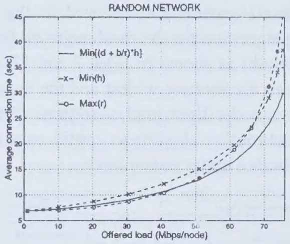


Fig. 9. Average connection time in the random network.

B. Optimization Functions

The results of the preceding section indicate that when the load is low and the network is not very regular, the shortest-widest algorithm can achieve a slightly better performance than the multi-cost routing algorithm that uses the optimization function $f(d,b,h) = \left(d+\frac{b}{r}\right) * h$. For the multi-cost routing algorithm to perform well even for irregular networks, we need to give more relative importance to the available capacity rather than to the number of links of a path. This can be achieved using a parametric function $f(d,b,h) = \left(d+\frac{b}{r}\right) * h^n$, where n is a real number in the interval (0,1) that reduces the importance of the number of hops. The value for n should be chosen based on the network topology; the smaller the degree of regularity in the network, the bigger the value of n should be.

Figure 10 plots the connection time versus the offered load for Mesh No. 3. The multi-cost routing algorithm with $n = 0.5$ clearly gives the best performance. Under light load conditions, the performance of the algorithm is similar to that of the shortest-widest algorithm, while for heavy load its performance is significantly better (see Fig. 10b). Although the new optimization function gives more importance to the available capacity r, it still considers the number of hops, and that is why its performance is good even under heavy load.

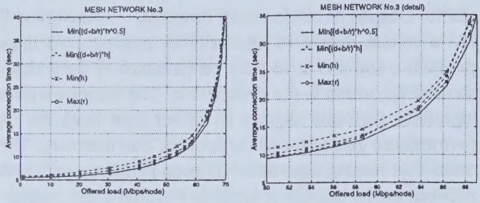


Fig. 10. (a) Average connection time in the mesh network No. 3. (b) Detail of the curves.

Similar improvements are achieved by the new multi-cost routing algorithm for the mesh network No. 2 and the random network, as shown in Figs. 7 and 12. The differences between the shortest-widest and the multi-cost routing algorithm under light load conditions disappear, and the multi-cost routing algorithms continues to outperform the other schemes under heavy load.

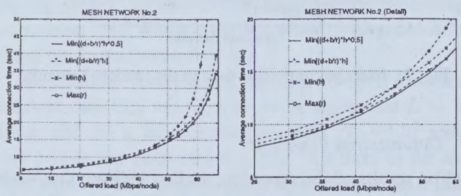


Fig. 11. (a) Average connection time in the mesh network No. 2. (b) Detail of the curves.

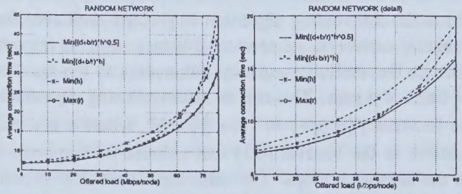


Fig. 12. (a) Average connection time in the random network. (b) Detail of the curves.

C. Specification of a Value for MCR

When a nonzero value for MCR is specified, a connection may not be accepted in the network. The efficiency of the routing algorithms can then be evaluated based on two measures: (a) the average connection time, which consists of the propagation delay, the transmission delay, and the total waiting time until the connection is accepted, and (b) the probability of success for a new connection.

We present simulation results for the mesh network No. 2 and the random network. We compare the widest-shortest, the shortest-widest and the multi-cost routing with minimization function $f = \left(d + \frac{b}{r}\right) * h^{0.5n}$. For the mesh No. 2 we use $MCR = 50$Mbps for traffic in the interval (5Mbytes, 500Mbytes), and $MCR = 5$Mbps for traffic in the interval (50Kbytes, 5Mbytes). For the random network, we use $MCR = 40$Mbps for traffic in the interval (5Mbytes, 500Mbytes), and $MCR = 4$Mbps for traffic in the interval (50Kbytes, 5Mbytes).

Figure 13 illustrates the connection time as a function of the offered load for Mesh No. 2, while Fig. 14 plots the success probability versus the offered load for the same network. The multi-cost routing algorithm with $n = 0.5$ gives the best results for both the average connection time and the success probability. The shortest-widest algorithm achieves good performance under light and medium load, but when the load increases, the average connection time suffers a considerable increase. (Both the figures for the connection time and the success probability show this decrease in efficiency at around 40Mbps). The widest-shortest algorithm gives acceptable results and a graceful performance degradation when the network load increases. Note that the average connection time reported in the figures includes the waiting time for retrials.

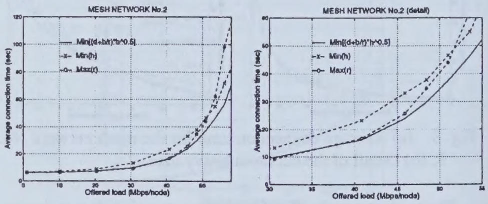


Fig. 13. (a) Average connection time in the mesh network No. 2. (b) Detail of the curves for medium and heavy load.

In Figs. 15a and 15b we illustrate the connection time and the success probability as a function of the offered load for the random network. The widest-shortest algorithm shows the worst results; the lack of regularity in the network makes this routing strategy inappropriate. The shortest-widest algorithm and the multi-cost routing

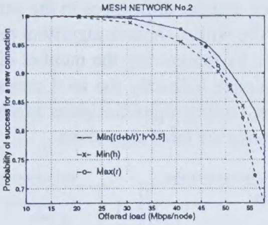


Fig. 14. Curves of the success probability for a new connection in the mesh network No. 2.

algorithm present similar results under light and medium load conditions. Further increases in the load, however, lead to a steep degradation in the performance of the shortest-widest algorithm.

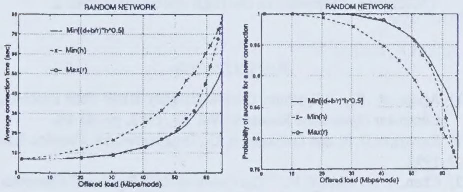


Fig. 15. (a) Average connection time in the random network. (b) Success probability in the random network.

D. *Influence of the Way the "Available Capacity" is Defined*

In this section we examine the way the definition of the "available capacity" affects the performance of the routing algorithm. We compare two ways of defining the available capacity: (a) As the estimate of the max-min rate that would be assigned to a new connection if accepted; this is the value that has been used in all the simulations. (b) As the unused capacity on each link; this definition does not reflect the nature of bandwidth sharing in max-min fair networks but it is a traditional way of defining the available capacity.

We compare the performance of the multi-cost routing algorithm with $n = 0.5$ using both definitions for the available capacity r. Figure 16a shows the results obtained for mesh No. 2, while Fig. 16b shows the results obtained for the random network. For light load, both definitions give similar results, because the paths with the most unused capacity are generally the paths that also give the biggest estimates for the max-min rate for a new connection. When the load increases, however, the definition that identifies available capacity with unused capacity leads to longer connection times, because it does not correctly reflect the nature of bandwidth sharing in max-min fair share networks.

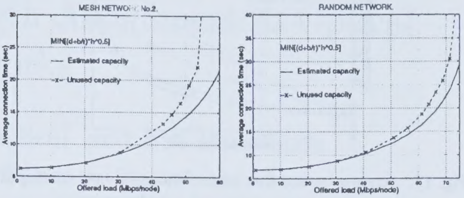


Fig. 16. (a) Average connection time in the mesh No. 2. (b) Average connection time in the random network.

E. *The Effects of Outdated Link Utilization Information*

The results presented in previous sections assume that the estimated rate for a new connection is obtained every time the cost of a path is calculated using the most updated link utilization information. These assumptions are clearly unrealistic. A more practical model would be to assume periodic updates of the link utilization information and the estimated rates. This would reduce the burden on the switches at the expense of a certain loss of efficiency in the routing.

In Fig. 17 we compare the the average connection time of the multi-cost routing algorithm with $n = 0.5$ for mesh No. 2 for two types of updating: 1) ideal instantaneous updating, and 2) periodic updating, with a period of $P = 10$, 50, or 100 seconds. Naturally, the algorithm that assumes ideal updating gives the best performance, and the larger the updating period is the worse are the results. Also, the larger the offered load is, the more severe is the effect of outdated information. The results obtained with $P = 10sec$ are very close to those obtained with ideal updating, which makes this choice of the period a practical alternative to ideal updating for the data sizes and link capacities we have assumed.

F. *Importance of the Propagation Delay*

In the preceding simulations, the link capacities allowed transmission times that were several orders of magnitude larger than the propagation delays. The situation is different for very high speed networks (with capacities of the order of Gbps), where the transmission times may be comparable to the propagation delays.

To study the effect of large propagation delays we consider again the mesh network No. 2, but assume link capacities that are 100 times larger, that is, 15.5 Gbps and 62.2 Gbps instead of 155Mbps and 622Mbps. We also

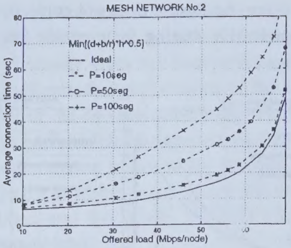


Fig. 17. Average connection time of the multi-cost routing algorithm using instantaneous and periodic updating.

increase the link lengths so that they take values in the interval (500Km, 300Km). The amounts of data transported by a session are again uniformly distributed in the intervals (50Kbytes, 5Mbytes) and (5Mbytes, 500Mbytes).

We compare the widest-shortest and the shortest-widest algorithms with the multi-cost routing algorithm that uses the minimization function $f = (d + \frac{b}{r}) * h$. We obtain the curves for the connection time separately for each category of traffic. Figure 18 illustrates the results obtained for the traffic in the interval (5Mbytes, 500Mbytes). Figure 18a, shows the results for the original mesh No. 2, and Figure 18b for the high speed mesh No. 2. We observe that the differences among the algorithms are proportionally the same in both networks: the transmission times are 100 times smaller in the high speed network, but this affects equally all the algorithms. The amount of data to be transmitted is still large enough to require, on the average, transmission times that are much larger than the propagation delays.

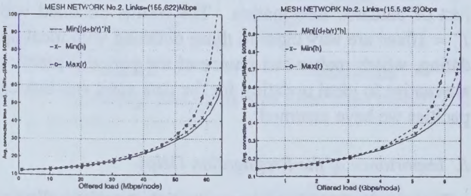


Fig. 18. (a) Average connection time for the traffic in the interval (5Mbytes, 500Mbytes) in the original mesh No. 2. (b) Average connection time for the traffic in the interval (5Mbytes, 500Mbytes) in the high speed mesh No. 2.

Figure 19 shows the results obtained for the traffic in the interval (50Kbytes, 5Mbytes). Figure 19a, illustrates the results for the original mesh No. 2, and Figure 19b for the high speed mesh No. 2. In the high speed network, the propagation delays are now comparable to the transmission times, and the differences in the performance are significant. The widest-shortest algorithm performs rather well, because by minimizing the number of hops, it also tends to minimize indirectly the total propagation delay. The multicost routing algorithm gives again the best performance among the three algorithms.

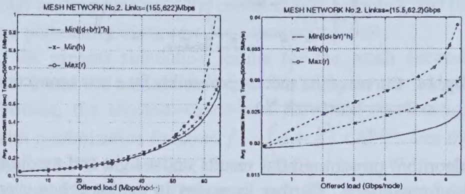


Fig. 19. (a) Average connection time for the traffic in the interval (50Kbytes, 5Mbytes) in the original mesh No. 2.(b) Average connection time for the traffic in the interval (50Kbytes, 5Mbytes) in the high speed mesh No. 2.

REFERENCES

[1] Ahuja, R .K., "Minimum Cost-Reliability Ratio Path Problem," Computer Operations Research, Vol. 15, 1988, pp. 83-89.

[2] Bertsekas D. P., and Gallager, R. G., Data Networks, Prentice-Hall, 1992.

[3] Chen, Y. L., Chin,Y. H., "The Quickest Path Problem," Computers Operations Research, Vol. 17, 1990, pp. 157-163.

[4] Chang, Y.-C., and Hsu, L.-H., "On Minimal Cost-Reliability Ratio Spanning Trees and Related Problems," Operations Research Letters, Vol. 19, 1996, pp. 65-69.

[5] Dijkstra, E. W., "A Note in two problems in connection with Graphs," Numerische Mathematik, Vol. 1, 1959, pp. 269-271.

[6] Fujimora, K., "Path Planning with Multiple Objectives," IEEE Robotics and Automation Magazine, 1996, pp. 33-38.

[7] Hung, Y. C., Chen, G. H., "The Quickest Path Problem," Proc. ICCI 91, LNCS 497, Springer-Verlag, 1991, pp. 44-46.

[8] Hansen, P., "Bicriterion Path Problems," Multiple Criteria Decision Making: Theory and Applications. Lecture notes in Economics and Mathematical Systems, 1980, pp. 109-127.

[9] Hansen, P., "A Bicriterion Shortest Path Algorithm," Europ. J. Operat. Res., Vol. 11, 1982, pp. 399-404.

[10] Jaffe, J. M. "Bottleneck Flow Control," IEEE Trans. on Commun., Vol. COM-29, No. 7, 1981, pp. 954-62.

[11] Kagaris, D., Pantziou, G. E., and Zaroliagis, C. D., "On the Computation of Fast Data Transmissions in Networks with Capacities and Delays, Algorithms and Data Structures, 1995, pp. 291-302.

[12] Martins, E. Q. V., "An Algorithm to Determine a Path with Minimal Cost / Capacity Ratio," Discrete Applied Mathematics, Vol. 8, 1984, pp. 189-194.

[13] Ma, Q., Steenkiste, P., Zang, H., "Routing High-Bandwidth Traffic in Max-Min Fair Share Networks," ACM SIGCOMM, 1996, pp. 206-217.

[14] Rosen, J. B., Sun, S. Z., . Xue, G. L., "Algorithms for the Quickest Path Problem and the Enumeration of Quickest Paths," Computers Operations Research, Vol. 6, 1991, pp. 579-584.

[15] Schwartz, M., and Stern, T. E., "Routing Techniques used in Computer Communication Networks," IEEE Trans. on Communications, Vol. 4, April 1980, pp. 539-552.

On Routing and Wavelength Assignment in Optical Networks

Andréa W. Richa[*], Arunabha Sen[†], and Bao Hong Shen
Department of Computer Science and Engineering
Arizona State University
Tempe, AZ 85287–5406, USA

Subir Bandyopadhyay
School of Computer Science
University of Windsor
Windsor, Ontario N9B 3P4, CANADA

Abstract

Routing and wavelength assignment in optical networks have been extensively studied in recent years. Most often the task is carried out in two separate phases: In the first phase, the routes are established between the specified source and destination nodes and in the second phase, wavelengths are assigned to the routes in such a way that the total number of distinct wavelengths used is minimized, subject to the constraint that no two lightpaths sharing a link are assigned the same wavelength. Although carrying out the task in two separate phases may not lead to an optimal solution, often this two phase approach is used since the problem of finding the paths that minimize the number of wavelengths is NP-complete (the wavelength assignment in the second phase alone is still NP-complete). During the route establishment phase, often *shortest-path* routes are used. Let λ_{sp} be the minimum number of wavelengths necessary to establish a set of p lightpaths between the source-destination pairs $\{(s_1, d_1), \ldots, (s_p, d_p)\}$ *if shortest-path routing* is used, and λ_{opt} is the optimum number of wavelengths necessary to establish the same set of lightpaths. The ratio $\lambda_{sp}/\lambda_{opt}$ can be as large as p (and p may be as large as $\Theta(n)$), as shown. An alternative approach to shortest-path routing is to use *minimum-load routes* in the route establishment phase. Let λ_{ml} be the minimum number of wavelengths necessary to establish a set of p lightpaths between the source-destination pairs $\{(s_1, d_1), \ldots, (s_p, d_p)\}$ *if minimum-load routing is used.* This paper shows that the ratio $\lambda_{ml}/\lambda_{opt}$ can be as large as $p/2$. For the minimum-load upper bound , we also show that p may be as large as $\Theta(\sqrt{n})$, implying that $\lambda_{ml}/\lambda_{opt}$ may be as large as $\Theta(\sqrt{n})$, where n is the number of nodes in the network. The bounds shown for $\lambda_{sp}/\lambda_{opt}$ and for $\lambda_{ml}/\lambda_{opt}$ are all tight.

---

[*] Supported in part by NSF CAREER Award CCR-9985284, NSF Grant CCR-9900304, and ASU Faculty-Grant-In-Aid CRR-L032.

[†] Corresponding author: Arunabha Sen, Telephone: 480-965-6153, Fax: 480-965-2751, e-mail: asen@asu.edu

1 Introduction

When routing in optical networks, one would like to assign wavelengths to the routing paths so as to use as few distinct wavelengths as possible. More specifically, given a set of p source-destination pairs $\{(s_1, d_1), \ldots, (s_p, d_p)\}$ in a graph $G = (V, E)$, one would like to find paths from s_i to d_i for $1 \leq i \leq p$, assigning respective wavelengths λ_i to the path between s_i and d_i, so that the total number of *distinct* wavelengths used is minimized, subject to the constraint that if two different paths share a link then they must use distinct wavelengths. Because of its importance, many researchers have studied this problem extensively in recent years (e.g., [2, 3, 4, 5, 13, 16, 17]). The problem of finding the optimal number of wavelengths for establishing the lightpaths between a set of source and destination nodes is NP-complete; even if the routing paths are given, the wavelength assignment of wavelengths to the paths alone is still NP-complete [3, 7, 13]. Thus, due to its inherent complexity, most often the problem is solved in two separate phases: In the first phase, the routes are established between the specified source and destination nodes; in the second phase, wavelengths are assigned to the routes so as to minimize the total number of distinct wavelengths used.

After the first phase, when the routes between the various source-destination pairs are established, a graph is constructed where each node represents a path and two nodes have an edge between them if the corresponding paths share a link. This graph is called the *edge-intersection graph* for the given routes. The *chromatic number* of this graph gives the *minimum number of wavelengths* needed if the routes established during the first phase are used for the lightpaths.

Although carrying out the task in two separate phases may not lead to an optimal solution, this two-phase approach is widely used because of its simplicity. During the route establishment phase, *shortest paths* are often used as the routes between the specified pairs of source and destination. Let λ_{sp} be the *minimum number of wavelengths* necessary to establish a set of p lightpaths between the source-destination pairs $\{(s_1, d_1), \ldots, (s_p, d_p)\}$ *if shortest-path routing* is used, and let λ_{opt} be the *optimum number of wavelengths* necessary to establish the a set of lightpaths between the given source-destination pairs. How far off from optimal, λ_{opt}, can λ_{sp} get ? We show that the solution obtained using shortest-path routing can be quite far off from the optimal: The ratio $\lambda_{sp}/\lambda_{opt}$ can be as large as p, and this upper bound is tight.

Shortest-path routing fares poorly when the *load* on a link (i.e. the number of lightpaths passing through the link) [24] is high. This situation can be avoided by distributing the lightpaths along different routes in such a way that the load on any single link is small. The *minimum-load routing algorithm* routes the paths in such a way that the maximum load of a link is minimized over all possible choices of routing paths.

A common belief is that the number of wavelengths required if we use minimum-load routes will be much smaller than the one required when shortest-path routing is used. The main contribution of this paper is to show that the minimum-load routes can perform almost as poorly as shortest-path routing, in terms of the number of wavelengths required to implement the chosen routing paths. Let λ_{ml} be the minimum number of wavelengths necessary to establish a set of p lightpaths between the source-destination pairs $\{(s_1, d_1), \ldots, (s_p, d_p)\}$ *if minimum-load routing* is used. We show that the ratio $\lambda_{ml}/\lambda_{opt}$ can be as high as $p/2$. This is a tight upper bound. In terms on n, the number of nodes in the network, we show that p may be as large as $\Theta(\sqrt{n})$, implying that $\lambda_{ml}/\lambda_{opt}$ may be as large as $\Theta(\sqrt{n})$.

The remainder of this paper is organized as follows. In Section 2 and 3, we prove tight upper bounds on the ratios $\lambda_{sp}/\lambda_{opt}$ and $\lambda_{ml}/\lambda_{opt}$, respectively. Finally, in Section 4, we make some concluding remarks, citing some important open problems.

2 Number of Wavelengths in Shortest-Path Routing

In this section we study the effect of shortest-path routing on the number of wavelengths required for path establishment. Shortest-path routing is the most frequently used routing algorithm for path establishment in an optical network [4, 23]. Although it is well known that shortest-path routing may not lead to the minimum number of wavelengths, it is often used in practice because of its simplicity and because of the advantages of having paths of short length between the source-destination pairs.

In this scheme of routing and wavelength assignment, once the shortest-path routes between the set of sources $S = \{s_1, \ldots, s_p\}$ and the set of destinations $D = \{d_1, \ldots, d_p\}$ are established, a graph is drawn where each node represents a path between s_i to d_i $(1 \leq i \leq p)$ and two nodes have an edge between them if the corresponding paths share a link. The *chromatic number* of this graph gives the minimum number of wavelengths needed *if the shortest-path routing is used for route establishment.* We will refer to this number as λ_{sp}.

Even if the chromatic number of the edge-intersection graph constructed above is computed exactly, this number does not necessarily give the fewest number of wavelengths necessary to establish the lightpaths between the corresponding sources in S and destinations in D. This is due to the fact that the number of wavelengths necessary is dependent on the routing algorithm used and it leaves open the possibility of some other routing algorithm establishing lightpaths between S and D which need fewer wavelengths.

We refer to the minimum number of wavelengths necessary to establish the lightpaths between S and D as λ_{opt}. In this section we ask the question: "How far off can λ_{sp} get from the optimal solution λ_{opt} ?". In other words, we want find the maximum possible value of the ratio $\lambda_{sp}/\lambda_{opt}$. Since $1 \leq \lambda_{opt} \leq \lambda_{sp} \leq p$, the ratio between λ_{sp} to λ_{opt} can be at most p.

We show in the following theorem that indeed the ratio between λ_{sp} to λ_{opt} can be as large as p. This scenario takes place if all the shortest paths are compelled to go through a single link in the network. In this case we say that the link is *heavily loaded.*

Theorem 1 *The ratio between λ_{sp} to λ_{opt} can be as large as p.*

Proof: Consider the graph shown in figure 1. The set of sources $s_1, \ldots, s_p$ and the set of destinations $d_1, \ldots d_p$ are shown in the figure. The shortest path from s_i to d_i for all $i, 1 \leq i \leq p$, is $s_i \rightarrow x \rightarrow y \rightarrow d_i$. If shortest-path routing is chosen, then all the paths will be compelled to pass through the link $x \rightarrow y$, making the load on that link to be p. Since the number of wavelengths must be at least as large as the load on the heaviest link, the number of wavelengths needed in this case must be p. However, if shortest-path routing is not used, then for each $i, 1 \leq i \leq p$, the paths between s_i to d_i could have been established as $s_i \rightarrow a_i \rightarrow b_i \rightarrow c_i \rightarrow d_i, 1 \leq i \leq p$. If routing is done in this manner, then only *one* wavelength will be needed to establish all the connections. In this case, $\lambda_{sp} = p$, $\lambda_{opt} = 1$, and the ratio $\lambda_{sp}/\lambda_{opt} = p$. ■

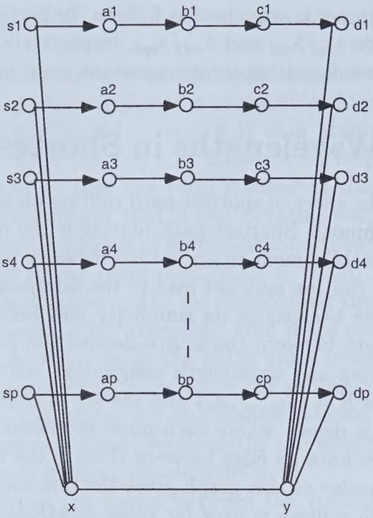


Figure 1:

3 Number of Wavelengths in Minimum-Load Routing

The shortest-path routing algorithm may heavily load any one link and as such skew the results of ratio $\lambda_{sp}/\lambda_{opt}$. As an alternate to the shortest-path routing, we consider minimum-load routing. In this scheme, the routes are established between the source and the destination nodes in such a way that the *load on a link* — e.g., the number of lightpaths passing through the link — is as small as possible.

Suppose that we need to establish a set of p lightpaths between the source-destination pairs $\{(s_1, d_1), \ldots, (s_p, d_p)\}$ in the graph $G = (V, E)$, where $|V| = n, |E| = m$. We want to establish the lightpaths in such a way that the number of distinct wavelengths used is minimum, subject to the constraint that no two lightpaths that share a link have the same wavelength. Suppose that $R_1, \ldots, R_k$ is the set of all possible ways of routing the traffic from the set of sources, $S = \{s_1, \ldots, s_p\}$ to the respective destinations in $D = \{d_1, \ldots, d_p\}$. Consider an $m \times k$ traffic matrix T whose (i, j)-th entry, $T[i, j]$, indicates the number of lightpaths through link e_i of the graph $G = (V, E)$ under the routing scheme R_j. Then the *absolute minimum load (AML)* on a link for the set of source-destination pairs $\{(s_1, d_1), \ldots, (s_p, d_p)\}$ is

$$AML((s_1, d_1), \ldots, (s_p, d_p)) = \min_{1 \leq j \leq k} [\max_{1 \leq i \leq m} T[i, j]]$$

We refer to a routing algorithm R_j that achieves the AML as a *minimum-load routing algorithm.*

Although the notion of AML is applicable to any networks, in case of optical networks it takes on additional significance. If *wavelength conversion* is allowed, the AML provides the exact number of wavelengths necessary to establish connection between the given source-destination pairs $\{(s_1, d_1), \ldots, (s_p, d_p)\}$. If wavelength conversion is not allowed, then the AML provides a *lower bound* on the number of wavelengths necessary to establish connection between the given source-destination pairs $\{(s_1, d_1), \ldots, (s_p, d_p)\}$. Figure 2 shows an example of a network where the number of wavelengths is larger than the AML, assuming that wavelength conversion is not allowed. Suppose that we need to establish paths between the nodes A to C, A to D, and C to D. In this case, all the three paths will be routed through the node B and none of the links will have to support more than *two* lightpaths, although the number of wavelengths necessary to establish these three connections will be *three.*

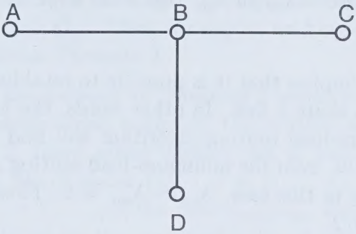


Figure 2:

The AML can be computed using techniques used for the *multicommodity flow* problem [8]. In a sense, it can be viewed as the dual of the multicommodity flow problem. In the multicommodity flow problem, the capacity of the links is known and the objective is to find the maximum flow that can be attained, subject to the link capacity constraints. In the AML computation, all source-destination pairs are known and the objective is to find the minimum link load that will be able to satisfy all the requests for the lightpaths. The mathematical programming formulation for the computation of the AML is given next.

The variable k is used to identify a lightpath from s_k to d_k, $(1 \leq k \leq p)$. A binary indicator variable $x_{i,j}^k$ is associated with each link (i, j) of the directed graph $G = (V, E)$. If the variable $x_{i,j}^k = 1$, it indicates that the link (i, j) is a part of the lightpath k on its journey from the source s_k to the destination d_k; if $x_{i,j}^k = 0$, then the link (i, j) is not part of such a path.

$\min C$

Subject to:

$$(i) \text{ for all } k, 1 \leq k \leq p, \quad \Sigma_{(i,j)\in E}\, x_{i,j}^k - \Sigma_{(j,\ell)\in E}\, x_{j,\ell}^k = \begin{cases} -1, & \text{if } j = s_k \\ 1, & \text{if } j = d_k \\ 0, & \text{otherwise} \end{cases}$$

$$(ii)\ \Sigma_{k=1}^{p}\, x_{i,j}^k \ \leq\ C,\ \forall\ (i,j) \in E$$

$$(iii)\ x_{i,j}^k \in \{0,1\},\ \forall (i,j) \in E$$

The multicommodity flow problem is NP-complete [7]. Approximation algorithms for these problem (with guaranteed performance bounds) are presented in [10, 22].

Once the minimum-load routes between the set of sources S and the set of destinations D are established, we construct the edge-intersection graph for these routes. The chromatic number of this graph gives minimum number of wavelengths needed *if the minimum-load routing is used for route establishment.* We will refer to this number as λ_{ml}. As before, we refer to the minimum number of wavelengths necessary to establish the lightpaths between S and D as λ_{opt}. In this section we answer the question "How far off can λ_{ml} get from the optimal solution λ_{opt}?". (In other words we want find the maximum possible value of the ratio $\lambda_{ml}/\lambda_{opt}$ in terms of p.) Since $1 \leq \lambda_{opt} \leq \lambda_{ml} \leq p$, the ratio between λ_{ml} to λ_{opt} can be at most p.

Theorem 2 *The ratio between λ_{ml} to λ_{opt} can be as large as $p/2$, and this upper bound is tight.*

Proof: If $\lambda_{opt} = 1$, then it implies that it is possible to establish the lightpaths in such a way that no two lightpath share a link. In other words, the load on any link is at most one. Clearly, the minimum-load routing algorithm will find these routes. Since these routes do not share any links, even the minimum-load routing algorithm will require only one wavelength. Therefore in this case, $\lambda_{ml} = \lambda_{opt} = 1$. Thus the ratio between λ_{ml} to λ_{opt} can only be at most $p/2$.

We now show that the ratio between λ_{ml} to λ_{opt} can indeed be as large as $p/2$. Consider the graph shown in Figure 3 (there we assume that $p = 4$, for simplicity). For each source-destination pair (s_i, d_i), $1 \leq i \leq p$, we construct a path P_i between s_i and d_i, following the "staircase" pattern as illustrated in the figure: Each path P_i intersects each other path P_j, where $1 \leq j \leq p$ and $i \neq j$, only at an edge $e(i,j)$ such that $e(i,j) \notin P_k$, for all $k \neq i, j$. Also, we construct alternative paths Q_i between s_i and d_i, which do not intersect any node or edge of the paths P_j but for the nodes in $S \cup D$: Each path Q_i consists of three edges $(s_i, x_{\lceil i/2 \rceil})$, $(x_{\lceil i/2 \rceil}, y_{\lceil i/2 \rceil})$, $(y_{\lceil i/2 \rceil}, d_i)$, where the nodes x_j's, y_j's, s_i's, and d_i's are all distinct.

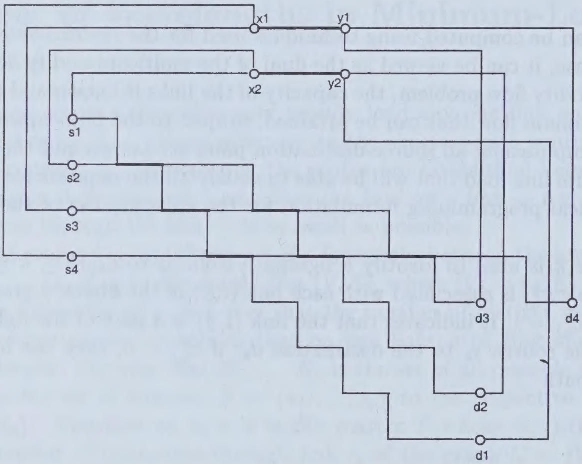


Figure 3:

A choice of paths between the pairs (s_i, d_i) which achieves minimum load 2 is $\{P_i : i = 1, \ldots, p\}$. The minimum wavelength number for this choice of paths is equal to p, since every path P_i intersects every other path P_j in the network. On the other hand, if routes were established through the paths Q_i, then the number of wavelengths required would be equal to 2. This is the minimum number of wavelengths required to establish paths between all the pairs (s_i, d_i) in the network — i.e., $\lambda_{opt} = 2$. Therefore $\lambda_{ml}/\lambda_{opt} = p/2$. ■

Corollary 3 *The ratio* $\lambda_{ml}/\lambda_{opt}$ *can be as large as* $\Theta(\sqrt{n})$.

Proof: We can construct a network with the properties outlined in the proof of Theorem 2 using $2p^2 + 2(p/2)$ nodes, since we need $2(p-1)$ edges per path P_i, and one pair of nodes x_j and y_j per paths Q_{2j-1} and Q_{2j}. Thus, for any n, we have $p = O(\sqrt{n})$ in this network and the corollary follows from Theorem 2. ■

The ratio $\lambda_{ml}/\lambda_{opt}$ in the network of Theorem 2 could vary significantly depending on the choice of the minimum-load paths by the routing algorithm. In the network constructed in Theorem 2, there exists another set of paths which also achieves the minimum load of 2: The set of paths $\{Q_i : i = 1, \ldots, n\}$. If these paths were used instead of the paths $\{P_i \; i = 1, \ldots, p\}$, the ratio $\lambda_{ml}/\lambda_{opt}$ would be equal to 1.

We present an example where the choice of a set of minimum-load paths does not affect the ratio $\lambda_{ml}/\lambda_{opt}$ significantly. For this, we modify the network in Theorem 2 by letting the paths Q_i be equal to $(s_i, x_{\lceil i/3 \rceil}), (x_{\lceil i/3 \rceil}, y_{\lceil i/3 \rceil}), (y_{\lceil i/3 \rceil}, d_i)$, where the nodes x_j's, y_j's, s_i's, and d_i's are all distinct: Now each edge (x_i, y_i) is used in exactly three paths Q_i. It is not hard to show that all the sets of paths with minimum load 2 can be characterized as follows: For $1 \leq k_i \leq 3$, choose k_i paths from $\{P_{3i-2}, P_{3i-1}, P_{3i}\}$, and choose the $(3 - k_i)$ remaining paths — connecting the yet uncovered pairs (s_j, d_j), for $3i - 2 \leq j \leq 3i$ — from $\{Q_{3i-2}, Q_{3i-1}, Q_{3i}\}$, for all $1 \leq i \leq n/3$. Then $p/3 \leq \lambda_{ml} \leq p$ in this network, with $\lambda_{ml} = p/3$ if we choose two Q_j's for every $1 \leq i \leq n/3$, and $\lambda_{ml} = p$ if we choose all the paths from the set $\{P_i : i = 1, \ldots, p\}$. Thus $p/9 \leq \lambda_{ml}/\lambda_{opt} \leq p/3$, implying that $\lambda_{ml}/\lambda_{opt} = \Theta(p)$, for any chioce of minimum-load paths in the network.

4 Conclusion

In this paper, we presented tight upper bounds on the ratios between λ_{ml} to λ_{opt}, and between λ_{sp} to λ_{opt}, showing that they are both as large as $\Theta(p)$. This shows that the common approaches of using shortest-path routes or minimum-load routes may lead to a wavelength assignment which is far from optimal. Thus, an important open question is to characterize the ratio $\lambda_{ml}/\lambda_{opt}$ for some specific classes of networks where a minimum-load algorithm may yield a better approximation for λ_{opt}. For example, it has been shown that for a two dimensional torus network, $\lambda_{ml} = \lambda_{opt}$ [12], if we consider the "full-connectivity" problem (i.e., every node is connected through a lightpath to every other node). It is also not hard to see that in a hypercube or butterfly network, equality is also achieved for full-connectivity. Are there any other classes for which λ_{opt} is not very far from λ_{ml}? For example, can anything be said about multi-mesh networks?

Another important question is to identify classes of networks for which the wavelength number can be approximated in polynomial time (computing the wavelength number remains NP-complete even if we consider restricted classes of networks as, for example, trees). Approximation algorithms have been proposed for many specific network classes: e.g., meshes [19, 18], trees [19, 9], rings [19], hypercubes [1].

References

[1] Y. Aumann, Y. Rabani. "Improved bounds for all optical routing", In *Proceedings of the 6th Annual Symposium on Discrete Algorithms*, pp 567-576, New York, NY, USA

[2] K. Bala, R.K. Chung and C.A. Brackett,"Optical Wavelength Routing, Translation, and Packet/Cell Switched Networks", *Journal of Lightwave Technology*, vol.14, no. 3, pp. 336-343, March 1996.

[3] D. Bannerjee and B. Mukherjee,"A practical approach for routing and wavelength assignment in large wavelength routed optical networks", *IEEE Journal on Selected Areas in communications*, vol. 14, no. 5, pp. 903-908, June 1996.

[4] S. Baroni and P. Bayvel,"Wavelength requirements in arbitrarily connected wavelength-routed optical networks", *IEEE Journal of Lightwave Technology*, vol. 15, no. 2, pp. 242-251, Feb. 1997.

[5] R.A. Barry and P.A. Humblet,"On the Number of Wavelengths and Switches in All-Optical Networks", In *IEEE Transactions on Communications*, vol. 42, no. 2/3/4, pp.583-591, Feb./Mar./April 1994.

[6] Y. Bartal and S. Leonardi, "On-line routing in all-optical networks", In *ICALP: Annual International Colloquium on Automata, Languages and Programming*, 1997.

[7] M.R. Garey and D.S. Johnson, Computers and Intractability: A Guide to the Theory of NP-Completeness. W.H. Freeman Publishers, San Francisco, 1978.

[8] T. C. Hu, Integer Programming and Network Flows, Addison Wesley Publishing Company, 1970.

[9] C. Kaklamanis, P. Persiano, "Efficient wavelength routing on directed fiber trees", In *ESA: Annual European Symposium on Algorithms*, Lecture Notes in Computer Science (1136), pp 460-470. Springer-Verlag, 1996

[10] P. Klein, S. Plotkin, C. Stein, E. Tardos, "Faster approximation algorithms for the unit capacity concurrent flow problem with applications to routing and finding sparse cuts", *SIAM Journal of Computing*, vol. 23, no. 3, pp. 466-487, 1994.

[11] J. Kleinberg, E. Tardos, " Disjoint Paths in Densely Encoded Graphs", In *36th Annual Symposium on Foundations of Computer Science*, pp 52-61, Milwaukee, Wisconsin, 23-25 October 1995.

[12] M. A. Marsan, A. Bianco, E. Leonardi, F. Neri, "Topologies for wavelength routing all optical-networks", In *IEEE/ACM Transactions on Networking*, 1(5):534-546, 1993

[13] A. Mokhtar and M. Azizoglu,"Adaptive wavelength routing in all-optical networks", *IEEE/ACM Transactions on Networking*, vol. 6, no. 2, pp. 197-206, April 1998.

[14] A. Mokhtar and M. Azizoglu,"Adaptive techniques for routing and wavelength assignment in all-optical WAN", *Proc. MWSCAS'96*, Ames, IA, pp. 220-225, 1996.

[15] B. Mukherjee, Optical Communication Networks, McGraw-Hill, 1997.

[16] N. Nagatsu, S. Okamoto and K. Sato,"Optical Path Cross-Connect System Scale Evaluation Using Path Accommodation Design for Restricted Wavelength Multiplexing", *IEEE Journal on Selected Areas in Communications*, vol. 14, no. 5, pp. 893-902, June 1996.

[17] R.K. Pankaj and R.G. Gallager,"Wavelength Requirements of All-Optical Networks", IEEE/ACM Transactions on Networking, vol. 3, no. 3, pp. 269-280, June 1995.

[18] Y. Rabani,"Path coloring on the mesh", In *37th Annual Symposium on Foundations of Computer Science*, pp. 400-409, Burlington, Vermont, 14-16 October, 1996.

[19] P. Raghavan, E. Upfal, "Efficient routing in all-optical networks". In *Proceedings of the Twenty-Sixth Annual ACM Symposium on Theory of Computing*, pp 134-143, Montreal, Quebec, Canada, 23-25 May, 1994.

[20] R. Ramaswami and K. Sivarajan, "Design of logical topologies for wavelength routed all optical network," *Proc. IEEE INFOCOM'95*, pp. 1316-1325, 1995

[21] R. Ramaswami and K. Sivarajan, Optical Networks: A Practical Perspective, Morgan Kaufmann Publishers Inc, 1998.

[22] F. Shahrokhi and D.W. Matula, "The maximum concurrent flow problem", *Journal of the Association of Computing Machinery*, vol. 37, no. 2, pp. 318-334, 1990.

[23] K. Sivarajan and R. Ramaswami, "Lightwave networks based on de Bruijn graphs," *IEEE/ACM Transactions on Networking*, vol. 2, no. 1, pp. 70-79, 1994.

[24] T.E. Stearn and K. Bala, Multiwavelength Optical Networks: A Layered Approach, Addison Wesley Longman Inc, 1999.

On the effects of switching in multi-fiber WDM networks*

Suresh Subramaniam[1], Hongsik Choi[2], and Hyeong-Ah Choi[3]
[1]Dept. of ECE, [3]Dept. of Computer Science
The George Washington University
Washington, DC 20052
{suresh, choi}@seas.gwu.edu

[2]Division of Computer Engineering
Hallym University
Chunchon, Korea.
hongsik@sun.hallym.ac.kr

Abstract

All-optical circuit-switched connections (lightpaths) can be established using optical channels in a multiwavelength wavelength routing network. Wavelength routing networks with multi-fiber links can utilize channels better through the switching capability between fibers. In this paper, we investigate the impact of switching capability by considering channel assignment for a given set of lightpaths in ring and tree topologies (off-line channel assignment). A set of lightpaths is characterized by a single parameter called load. First, we quantify the effects of switch size by deriving a bound on the load as a function of the switch size in ring topologies. In tree topologies, we present an example to show that switching between fibers is not sufficient, in general, to obtain a channel assignment using L channels per link where L is the load.

1 Introduction

WDM wavelength routing networks enable the formation of lightpaths which are envisioned to be the transport links for a variety of networks that may be overlaid on the physical optical infrastructure. Lightpaths are all-optical circuit-switched paths which are formed by choosing one channel (wavelength) on each link from the source to the destination of the lightpath and concatenating them. The channels that can be chosen for a lightpath depend on the capability of the nodes to perform wavelength conversion, i.e., switching an incoming channel at one wavelength to an outgoing channel at a different wavelength. If the nodes are incapable of wavelength conversion, then all channels of a lightpath have to be on a single wavelength. Wavelength conversion, by allowing more flexibility in assigning channels to a lightpath, is likely to improve channel utilization.

Practical optical networks of the future are likely to have many strands of fiber bundled together to form a single link. The number of channels available on each link in

---

*This work was supported in part by the NSA under LUCITE contract #MDA904-98-C-A081, and the NSF under grant ANI-9973098.

such a multi-fiber network is equal to the number of fibers multiplied by the number of wavelengths that each fiber can accommodate. Channel utilization can be increased even without wavelength conversion in multi-fiber networks if switching between channels at the same wavelength is enabled. For example, space switches may be used to switch an incoming channel to an outgoing channel at the same wavelength, but on an arbitrary fiber. In this case, lightpaths may be formed by concatenating channels on any one of the many fibers on each link along the lighpath's route, but on a single wavelength. Note that channels at different wavelengths may not be used if wavelength conversion is not available.

In this paper, our focus is on multi-fiber networks with some given switching capability between fibers, but without wavelength conversion capability. Our motivation for considering such networks is the following. Based on the architectures presented in [1], a wavelength router with K links per node, d fibers per link, and W wavelengths per fiber that allows arbitrary switching between fibers has Kd mux/demuxes of size W, and W switches of size $Kd \times Kd$. To achieve the same channel switching capability with single-fiber links (and correspondingly more wavelengths per fiber) and using limited wavelength conversion requires fewer (K) mux/demuxes and KdW wavelength converters.[1] Given the expense of all-optical wavelength converters, a network architecture using multi-fiber links with some inter-fiber switching appears to be an attractive alternative. We will not consider networks with wavelength conversion in this paper.

1.1 Network Model

The network consists of N nodes numbered $0, 1, \ldots, N-1$, and is denoted by a graph $G = (V, E)$. Each directed link $e \in E$ consists of d fibers, numbered $0, 1, \ldots, d-1$, where all fibers are oriented in the direction of the link. Each fiber is assumed to carry W wavelengths, numbered $0, 1, \ldots, W-1$. A channel in the network is identified by its link, fiber, and wavelength – the channel on link $e \in E$ on wavelength i and fiber j is denoted by $\lambda_i^j(e)$. We consider bidirectional network topologies in this paper, i.e., the presence of a link (v_i, v_j) implies the presence of link (v_j, v_i).

Two channels are said to be *attached* at a node if they may be connected together at the node to form a lightpath. For example, attached channels are joined by lines within the node in Figure 1. The model in this paper allows channels to be attached only if they are on the same wavelength. The inter-fiber switching capability is determined by the *switching degree* $c \leq d$. A switching degree of c implies that the d incoming channels and the d outgoing channels at a given wavelength on an I/O link pair can be partitioned into $\lceil \frac{d}{c} \rceil$ sets each such that: (a) no set has more than c channels, and (b) all channels in each input (output) set are attached to all channels in exactly one output (input) set. A node with switching degree c can be implemented with $c \times c$ switches. For example, the channels on two wavelengths of an I/O link pair with $d = 4$ are shown in Figure 1. The switching degree of this node is $c = 2$.

In this paper, we investigate the effects of inter-fiber switching by restricting ourselves to the off-line channel assignment problem in ring and tree topologies. In off-line channel assignment, a set of lightpaths is given, and a channel assignment (defined below) is obtained by considering all the lightpaths at once. By contrast, on-line channel assignment requires channels to be assigned to lightpaths that arrive and depart dynamically. Besides providing insight into and performance bounds for the more difficult on-line case,

---

[1] A wavelength converter here converts an arbitrary wavelength to a fixed wavelength.

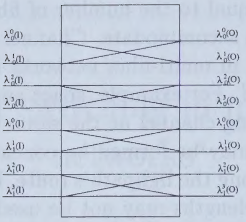


Figure 1: Channels attached at a node.

off-line channel assignment is useful in bandwidth provisioning applications.

A lightpath set is defined by a set of directed routes[2] on the network called a *request*. Given a request, a *channel assignment* assigns, for every route in the request, a channel on each link of the route such that the channels are attached at the intermediate nodes. We say that a request can be *established* if a channel assignment exists for the request. A request is assumed to be made up of bidirectional lightpaths, i.e., a lightpath between v_i and v_j implies one between v_j and v_i. We also assume that the set of channels in each direction used by a lightpath is the same in the two directions except that channels are on oppositely directed links. Because of this assumption, we only consider links and lightpaths directed one way (clockwise in rings and away from the root in trees). For any request R, let $L_{\min}(R)$ and $L_{\max}(R)$ be the minimum (maximum, respectively) number of routes passing over any link. The request is characterized by the single parameter called the *load* which is defined to be $L_{\max}(R)$. A request R is called *uniform* if $L_{\min}(R) = L_{\max}(R)$, and the load is denoted by $L(R)$. When the context is clear, we simply denote the load by L.

The rest of the paper is organized as follows. Related previous work is reviewed in Section 1.2. In Section 2, a bound on the load of a request that can be established in a bidirectional ring topology is obtained. Tree topologies are considered in Section 3 where we present an example tree and request to show the impossibility of obtaining a channel assignment using L channels per link in multi-fiber trees with inter-fiber switching, but without wavelength conversion. Finally, the paper is concluded in Section 4.

1.2 Related Previous Work

From the discussion in Section 1, it is clear that there is a logical relationship between multi-fiber networks with inter-fiber switching and single-fiber networks with limited wavelength conversion, with respect to channel assignment. Therefore, related previous work on the channel assignment problem in multi-fiber networks includes work on limited wavelength conversion networks as well.

Multi-fiber wavelength-routing optical networks have been studied under statistical dynamic traffic in [2, 3, 4, 5, 6]. Off-line channel assignment in single-fiber rings and trees with limited wavelength conversion is presented in [7]. In that work, the amount of wavelength conversion was characterized by a parameter called the conversion degree. A node with conversion degree k allows each channel to be switched to at most k other channels. Ring networks with fixed wavelength conversion configurations at one node that support a load of $W-1$, and those with a limited amount of wavelength conversion that support a load of W were presented in [7]. The work in the first part of this paper is different from the work of [7] in that the results obtained here are under more restricted channel connectivity. For example, the channel connectivity patterns at a multi-fiber router and a single-fiber router with the same number of channels are shown in Figures 2(a) and (b), respectively. In Figure 2(a), there are two fibers per link, two wavelengths per fiber, and the switching degree is two. In Figure 2(b), there are four wavelengths, and the conversion degree is two. Note that the connectivity of Figure 2(b) is not achievable with switching degree two in the multi-fiber node because wavelength conversion is absent. Therefore, more channels (but with possibly less overall cost) may be required to establish the same request in a multi-fiber network with switching degree k than in a single-fiber network with conversion degree k.

---

[2] Note that we assume shortest paths in bidirectional rings.

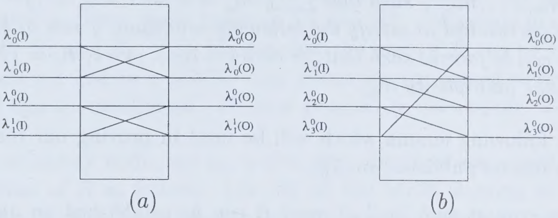


Figure 2: Channel connectivities at a node with (a) switching degree 2, and (b) conversion degree 2.

Bounds on the number of wavelengths required to establish a request in d-fiber rings with $d \times d$ switches at every node are presented in [8] for a variety of constraints on lightpath lengths. A result we obtain here is a generalization of one of their results. Channel assignment in single-fiber trees was studied in [9] where an upper bound of $\lfloor \frac{3L}{2} \rfloor$ for the number of wavelengths was established using the well-known Shannon bound on edge-coloring of multigraphs [10].

2 Ring Topologies

In this section, we obtain a bound for the load of an arbitrary uniform request that can be established, as a function of the number of fibers d, the number of wavelengths W, and the switching degree c, in a ring topology with N nodes.

We start by presenting some terminology, part of which was introduced in [7].

Definition 2.1 *A* multicycle of channels *(MCC) is a sequence of distinct channels* $(c_0, c_1, \ldots, c_{l-1})$, *where l is a multiple of N, that starts at some node u, goes clockwise around the ring one or more times, and ends at node u. Moreover, the channels c_i and $c_{i+1} \bmod l$ are attached for $i = 0, \ldots, l-1$. The* multiplicity *of an MCC is the number of times it goes around the ring, i.e., l/N.*

Definition 2.2 *A* multicycle of routes *(MCR) is a sequence of routes* $(r_0, r_1, \ldots, r_{l-1})$ *for some l such that $d_i = s_{i+1} \bmod l$ for $i = 0, \ldots, l-1$, where s_i and d_i are the source and destination nodes of route r_i. In other words, an MCR is a sequence of routes that starts at a node, goes clockwise around the ring one or more times, and ends at the same node. The* multiplicity *of an MCR is the number of times the ring is circumvented.*

Any uniform request can be partitioned into a sequence of MCRs such that each route in the request is in exactly one MCR and every route is in some MCR. To find an MCR partition of a request, form an MCR starting (and ending) at an arbitrary node s_0. Then, form another MCR using the remaining routes in the request starting at an arbitrary node s_1, and so on until all routes in the request are in some MCR. This is possible because the request is uniform, i.e., if there is a route starting at a node, then there must be one ending at that node. Notice that the sum of the multiplicities of the MCRs is equal to the load of the request. Conversely, a set of MCRs forms a uniform request whose load is equal to the sum of the multiplicities of the MCRs.

Definition 2.3 *For an arbitrary positive integer β, a ring with α channels per link is called a β-MCC ring network, for $\alpha = j\beta$ where $j \geq 1$ is an integer, if, given arbitrary positive integers $m_{i_0}, \ldots, m_{i_{l_i-1}}$ such that $\sum_{n=0}^{l_i-1} m_{i_n} = \beta$ for $i = 0, \ldots, j-1$, the channels of the ring can be partitioned to satisfy the following condition. j sets of MCC collections $\mathcal{H}_0, \mathcal{H}_1, \ldots, \mathcal{H}_{j-1}$ can be formed such that for each $i = 0, \ldots, j-1$, $\mathcal{H}_i = \{H_{i_0}, \ldots, H_{i_{l_i-1}}\}$, where MCC H_{i_n} has multiplicity m_{i_n}.*

We prove the following lemma which will be used in proving our main result. The proof is based on results published in [7].

Lemma 2.1 *Any request with load at most β can be established on any β-MCC ring network.*
Proof: Without loss of generality, consider a request R with $L(R) = \beta$. Let an MCR partition of the request be $\{M_0, \ldots, M_{l-1}\}$ with multiplicities $m_0, \ldots, m_{l-1}$, respectively, where $\sum_{i=0}^{l-1} m_i = L(R) = \beta$. Since the ring is β-MCC, there exists a set of MCCs $H_0, \ldots, H_{l-1}$ such that H_i has multiplicity m_i for $i = 0, \ldots, l-1$. The channel assignment is obtained by simply assigning the channels of H_i to the routes in M_i. ■

Our main result in this section is the following.

Theorem 2.1 *Consider a d-fiber ring with W wavelengths per fiber, and suppose one arbitrary node has switching degree c where $d = kc$, and all other nodes have switching degree one (incoming channels are attached to corresponding outgoing channels). In this ring network, any request that uses shortest path routing and whose load L satisfies the following condition can be established.*

$$L \leq \begin{cases} kW(2c-1)/2 & kW \text{ even,} \\ (kW(2c-1)+1)/2 & kW \text{ odd.} \end{cases}$$

Before proving the theorem, we prove the following lemma.

Lemma 2.2 *The ring network described in Theorem 2.1 is a c-MCC ring.*
Proof: We only give an outline of the proof here. For simplicity, we assume that node 0 has switching degree c and this is achieved using k switches of size $c \times c$. Moreover, we assume that at node 0, for every $i = 0, \ldots, W-1$, $\lambda_i^{jc}(I), \ldots, \lambda_i^{(j+1)c-1}(I)$ are all attached to every one of $\lambda_i^{jc}(O), \ldots, \lambda_i^{(j+1)c-1}(O)$, for every $j = 0, \ldots, k-1$, where I and O denote the input and output links of node 0. The channels can then be partitioned into kW sets of MCC collections $\{\mathcal{H}_0, \ldots, \mathcal{H}_{kW-1}\}$ as follows. For every $i = 0, \ldots, kW-1$, let $w \stackrel{\text{def}}{=} \lfloor i/k \rfloor$ and $f \stackrel{\text{def}}{=} c(i \bmod k)$, and define $\mathcal{H}_i \stackrel{\text{def}}{=} \{\lambda_w^j(e) : j = f, \ldots f+c-1; e \in E\}$. Now, consider an arbitrary set of MCCs from the partition above, $\mathcal{H}_i$, and suppose we

are given arbitrary positive integers $m_{i_0}, m_{i_1}, m_{i_{l-1}}$ such that $\sum_{n=0}^{l-1} m_{i_n} = c$. Then, the set of channels $\{\lambda_w^f(e), \ldots, \lambda_w^{f+m_{i_0}-1}(e) : e \in E\}$ form an MCC H_{i_0} with multiplicity m_{i_0} because $\lambda_w^{f+j}((N-1,0))$ and $\lambda_w^{f+(j+1)\mathrm{mod} m_{i_0}}((0,1))$ are attached for $j = 0, \ldots, m_{i_0} - 1$. This is because node 0 has switching degree c. Also note that $\lambda_w^j((i, i+1)$ is attached to $\lambda_w^j((i+1, (i+2) \bmod N))$ for $i = 0, \ldots, N-2$. The MCCs $H_{i_1}, \ldots, H_{i_{l-1}}$ are similarly formed. Therefore, the ring is a c-MCC ring. ∎

Proof of Theorem 2.1: The cases of kW even and kW odd are considered separately. **Case 1:** Let kW be even. Let R be any uniform request with load $L^* \overset{\text{def}}{=} kW(2c-1)/2$. We will decompose R into three uniform requests R_1, R_2, and R_3 such that every route in R is present in exactly one of the three sets. Besides the routes in R, the requests may contain some dummy routes. We will also show that $L(R_1) = L(R_2) = c$ and $L(R_3) = L^* - 2c + 1$. Then, from Lemmas 2.1 and 2.2, c channels per link each can be used to establish R_1 and R_2. This implies that the load of R can be reduced by $2c - 1$ using $2c$ channels per link in a c-MCC ring. Since the number of channels available per link is kcW, R can be established. We now proceed to decompose R into R_1, R_2, and R_3.

Start at an arbitrary node, say u_0, which is an origin of some route in R, and form an MCR partition of R as follows. Let M_0 be the MCR starting at node u_0. Then, choose the next clockwise node around the ring, say u_1, which is the origin of a route in $R - M_0$, and form an MCR M_1. Let $M_0, \ldots, M_{l-1}$ be the MCR partition formed in this way, and let $m_0, \ldots, m_{l-1}$ be their multiplicities. Without loss of generality, we may assume $u_0 < u_1 < \ldots < u_{l-1}$.

Let j_1 be the smallest integer such that $\sum_{i=0}^{j_1} m_i \geq c$, and let $\xi \overset{\text{def}}{=} c - \sum_{i=0}^{j_1-1} m_i$, and $\theta \overset{\text{def}}{=} \sum_{i=0}^{j_1} m_i - c = m_{j_1} - \xi$. Form a request R' starting from u_{j_1} (using the routes in M_{j_1}) until a route $r_1 \in M_{j_1}$ such that r_1 crosses u_{j_1} for the ξth time (among routes in M_{j_1}) or ends at u_{j_1} after crossing it $\xi - 1$ times. Such a route must exist by definition of j_1 and ξ. Here, R' is formed by considering routes in sequence where the destination of one route is the origin of the next, proceeding clockwise. If r_1 ends at u_{j_1}, we set $s_1 \overset{\text{def}}{=} u_{j_1}$, and include r_1 in R'. Otherwise, let s_1 be the origin of r_1, and r_1 is not included in R'. Let $d_0 = (s_1, u_{j_1})$ be a dummy path (of length zero if $s_1 = u_{j_1}$). Now, $X \overset{\text{def}}{=} R' \cup \{d_0\}$ is an MCR of multiplicity ξ, and therefore, $R_1 \overset{\text{def}}{=} \cup_{i=0}^{j_1-1} M_i \cup X$ is a uniform request with load c.

Now, let j_2 be the smallest integer such that $\sum_{i=0}^{j_2} m_i \geq 2c - 1$. The following two cases arise: (a) $j_2 = j_1$ and (b) $j_2 > j_1$.

Case (a): $j_2 = j_1 \Rightarrow \theta = \sum_{i=0}^{j_1} m_i - c \geq c - 1 \Rightarrow m_{j_1} - \xi \geq c - 1$. Starting from s_1, form a request R'' (using routes in M_{j_1} and proceeding clockwise as for R') until $r_2 \in M_{j_1}$ which crosses s_1 for the cth time, or ends at s_1 after crossing s_1 $(c-1)$ times, or ends at u_{j_1} after crossing s_1 $(c-1)$ times is seen. Such a route must exist because $L_{\max}(R_1 \cap M_{j_1}) = \xi$ and because $m_{j_1} - \xi \geq c - 1$. If r_2 ends at s_1 (correspondingly, u_{j_1}), let $s_2 \overset{\text{def}}{=} s_1$ (correspondingly, u_{j_1}), and include r_2 in R''. Otherwise, r_2 crosses s_1, and we let s_2 be the origin of r_2, and do not include r_2 in R''. Now, let $d_2 \overset{\text{def}}{=} (s_2, s_1)$ be a dummy path (of zero length if $s_2 = s_1$). Then, it is clear that $R_2 \overset{\text{def}}{=} R'' \cup \{d_2\}$ is an MCR of multiplicity c, and hence, has load c. We now claim that d_0 and d_2 do not overlap. To see this, note that d_0 starts at s_1 and d_2 ends at s_1. Let $\mathrm{len}(r)$ denote the length of route r. Then, if d_0 and d_2 overlap, $\mathrm{len}(d_0) + \mathrm{len}(d_2) > N \Rightarrow \mathrm{len}(r_1) + \mathrm{len}(r_2) > N$, contradicting the fact that shortest path routing is used. Now, let $d_3 \overset{\text{def}}{=} (u_{j_1}, s_2)$. Then, $R_3 \overset{\text{def}}{=} R - (R_1 \cup R_2) \cup \{d_3\}$ is a uniform request. Since R_1 and R_2 have load c, and since d_0 and d_2 do not overlap, $L_{\min}((R_1 \cup R_2) \cap R) = 2c - 1$, and therefore, $L(R_3) = L^* - 2c + 1$.

Case (b): $j_2 > j_1 \Rightarrow \theta < c - 1$. Let $x \overset{\text{def}}{=} \sum_{i=j_1+1}^{j_2-1} m_i$, and let $\psi \overset{\text{def}}{=} c - x - \theta - 1$. Starting from u_{j_2}, form a request R'' (using routes in M_{j_2}) and proceeding clockwise as before) until $r_2 \in M_{j_2}$ which crosses s_1 for the ψth time or ends at s_1 after crossing $(\psi - 1)$ times is seen. Such a path must exist because $m_{j_2} \geq 2c - 1 - \sum_{i=0}^{j_2-1} m_i = 2c - 1 - (\sum_{i=0}^{j_1} m_i + \sum_{i=j_1+1}^{j_2-1} m_i) = 2c - 1 - (\theta + c + x) = c - x - \theta - 1 = \psi$. In the former case, define s_2 to be the origin of r_2, and in the latter case, $s_2 \overset{\text{def}}{=} s_1$. Let $d_1 \overset{\text{def}}{=} (u_{j_1}, u_{j_2})$, and let $d_2 \overset{\text{def}}{=} (s_2, s_1)$. Now, $X \overset{\text{def}}{=} (M_{j_1} - R_1) \cup \{d_1\} \cup R'' \cup \{d_2\}$ is an MCR of multiplicity $c - x$. Therefore, $R_2 \overset{\text{def}}{=} \cup_{i=j_1+1}^{j_2-1} M_i \cup X$ is a request with load c. Now, $R_3 \overset{\text{def}}{=} R - (R_1 \cup R_2) \cup \{d_3\}$ where $d_3 \overset{\text{def}}{=} (u_{j_2}, s_2)$ is a uniform request. To prove that $L(R_3) = L^* - 2c + 1$, we must show that d_0, d_1, and d_2 do not overlap. We proceed to show this next.

Recall that $d_0 = (s_1, u_{j_1})$, $d_1 = (u_{j_1}, u_{j_2})$, and $d_2 = (s_2, s_1)$. We will prove that u_{j_1}, u_{j_2}, and s_2 (if $s_2 \neq s_1$) are visited in this order in a clockwise traversal of the ring starting from s_1. First, suppose u_{j_2} lies between s_1 and u_{j_1}. Then, there would be a route ending (and, therefore, starting because R is uniform) between u_{j_1} and s_1 because of shortest path routing. Let the starting route be $r_3 \in M_{j_2}$. However, because of the way the MCR partition was constructed, (i.e., after the ith MCR ends, the $(i+1)$th MCR starts at the nearest possible clockwise node for $i = 0, 1, \ldots, l - 2$), r_3 must have been included in one of $M_0, \ldots, M_{j_2-1}$, a contradiction.

Now, suppose s_2 lies between s_1 and u_{j_1}. Then $\text{len}(r_2) > \lfloor N/2 \rfloor$, contradicting the shortest path assumption. Finally, suppose s_2 lies between u_{j_1} and u_{j_2}. Then, there would be a route $r_4 \in M_{j_2}$ starting between u_{j_1} and u_{j_2}, which contradicts the fact all such routes must have been included in one of $M_0, \ldots, M_{j_2-1}$. Therefore, d_0, d_1, and d_2 do not overlap, and since R_1 and R_2 have load c, $L_{\min}((R_1 \cup R_2) \cap R) = 2c - 1$, and therefore, $L(R_3) = L^* - 2c + 1$.

Case 2: Let kW be odd. Suppose a request with load $L^* = (kW(2c-1)+1)/2$ is given. Note that $(kW(2c-1)+1)/2 = (kW-1)(2c-1)/2 + c$. Since $(kW - 1)$ is even, from Case 1, L^* can be reduced by $((kW-1)(2c-1)/2$ using $(kW-1)c$ channels. Then, the remaining c channels can be used to establish the remaining request of load c, since the ring is c-MCC. ∎

3 Wavelength Assignment in Tree Topologies

A lower bound for the number of channels per link required to establish a uniform request of load L in any topology is L. In single-fiber trees, an upper bound of $\lfloor 3L/2 \rfloor$ for the number of channels per link was presented in [9]. Clearly, this upper bound also holds for multi-fiber trees with inter-fiber switching. Here, we address the following question: In a multi-fiber tree network with $W > 1$ wavelengths per link, is it possible to establish any request with load L using L channels per link and inter-fiber switching alone (when $d < L$; otherwise, it is obviously possible)? We answer the question in the negative by presenting an example below.

Suppose that a request R with load L^* can be established using L^* channels per link and nodes of switching degree d, $1 \leq d < L^*$. Then there must be a uniform request $Q_d \subset R$ such that $L(Q_d) = d$. For example, Q_d consists of all the routes assigned channels on a particular wavelength. We prove that this is not always possible by constructing an example.

Our main result in this section is:

Theorem 3.1 *There exists a binary tree T and a request R with $L(R) = 2^t$ for an integer $t > 0$, such that, there is no uniform request $Q_d \subset R$ with $L(Q_d) = d$, for $1 \leq d < L(R)$.*

In the following, we proceed to prove the theorem using several lemmas that we prove along the way. The next three technical lemmas are properties of integers. First, note that any positive integer n can be represented as a multiple of two integers, $n = pq$, where $p = 2^k$ and $q = 2j + 1$ for nonnegative integers k and j. This is because of the fact that any positive integer has a unique prime factorization. Given a positive integer $n = 2^k(2j+1)$, we recursively define a sequence $(m_0, m_1, \ldots, m_k)$ as follows: $m_0 = n$; and $m_i = n - m_{i-1}/2$ for $1 \leq i \leq k$. We now establish the following results for $(m_0, \ldots, m_k)$.

Lemma 3.1 *$m_i > 0$ for $0 \leq i \leq k$ and $m_i < n$ for $1 \leq i \leq k$.*

Proof: The proof is by induction on i. Clearly, $m_0 > 0$, $m_1 > 0$, and $m_1 < n$. Assume that $0 < m_i < n$ for any i in $1 \leq i < k$ and consider $m_{i+1} = n - m_i/2$. By the induction hypothesis, $0 < m_i < n$, and this implies that $0 < m_{i+1} < n$. ∎

Lemma 3.2 *For each $i > 0$, there exists a positive integer x_i such that $m_i = n(2x_i - 1)/2^i$.*

Proof: The proof is by induction on i. For $i = 0, 1$, setting $x_0 = x_1 = 1$ proves the claim. Assume there exists such an x_i for each i, $1 \leq i < k$, and consider m_{i+1}. Note that $m_{i+1} = n - m_i/2$, and by the induction hypothesis, we have $m_{i+1} = n - n(2x_i - 1)/2^{i+1} = n(2(2^i - x_i + 1) - 1)/2^{i+1}$. By setting $x_{i+1} = 2^i - x_i + 1$, the claim is proved for $i + 1$. ∎

Lemma 3.3 *There exists a non-negative integer y_i for each i in $0 \leq i \leq k$ such that $m_i = 2y_i$ for $0 \leq i < k$ and $m_k = 2y_k + 1$.*

Proof: Suppose $k = 0$, i.e., n is odd. Then, $m_0 = n = 2j + 1$, and setting $y_0 = j$ proves the lemma. Now, suppose $k > 0$. For $i = 0$, setting $y_0 = 2^{k-1}(2j + 1)$ gives $m_0 = 2y_0$ which proves the lemma. Let $0 < i < k$. By Lemma 3.2, there exists a positive integer x_i for each i such that $m_i = 2^{k-i}(2x_i - 1)(2j + 1)$. Hence, by setting $y_i = 2^{k-i-1}(2x_i - 1)(2j + 1)$, the lemma is proved for $0 < i < k$. Finally, consider $i = k$. By Lemma 3.2, there exists a positive integer x_k such that $m_k = (2x_k - 1)(2j+1) = 2(2jx_k + x_k - j - 1) + 1$. It is easy to see that $2jx_k + x_k - j - 1 \geq 0$; hence, setting $y_k = 2jx_k + x_k - j - 1$ completes the proof of the lemma. ∎

From Lemmas 3.1 and 3.3, we have the following result.

Corollary 3.1 *m_i is positive, even for each i in $0 \leq i < k$ and m_k is positive, odd.*

We now construct a binary tree T and a uniform request R with load $L^* = 2^t$ for an integer t, as follows. We let $n = 2^t$, and define the sequence of $(m_0, \ldots, m_t)$ as described, i.e., $m_0 = 2^t$, and $m_i = 2^t - m_{i-1}/2$ for $1 \leq i \leq t$. From Corollary 3.1, we know that m_i is a positive even integer for $0 \leq i < t$, and that m_t is a positive odd integer.

The reader is referred to Figure 3 in following the construction presented below. The vertex set $V(T)$ is defined as $V(T) \overset{\text{def}}{=} \{u_i \mid 0 \leq i \leq t+1\} \cup \{v_i \mid 0 \leq i \leq t\}$, and the edge set $E(T)$ is defined as $E(T) \overset{\text{def}}{=} \{(u_i, u_{i+1}) \mid 0 \leq i \leq t\} \cup \{(u_i, v_i) \mid 0 \leq i \leq t\}$.

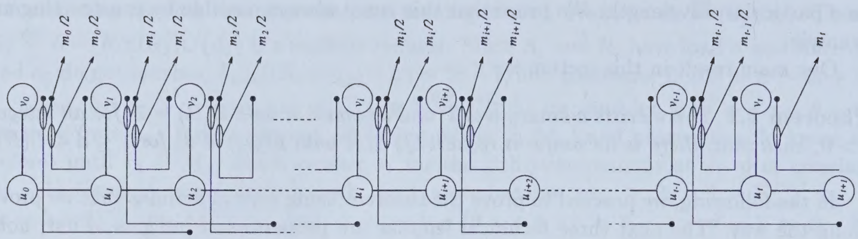


Figure 3: An example binary tree.

A uniform request R is defined as $R \overset{\text{def}}{=} \cup_{i=0}^{t}(R_i^1 \cup R_i^2)$, where (a) for $0 \le i < t$, R_i^1 denotes the set of $m_i/2$ routes, each connecting v_i to u_{i+2} through u_i and u_{i+1}, i.e., $\langle v_i, u_i, u_{i+1}, u_{i+2}\rangle \in R_i^1$, and R_i^2 denotes the set of $m_i/2$ routes each connecting v_i to v_{i+1} through u_i and u_{i+1}, i.e., $\langle v_i, u_i, u_{i+1}, v_{i+1}\rangle \in R_i^2$, and (b) R_t^1 denotes the set of m_t routes each connecting v_t to u_{t+1} through u_t, i.e., $\langle v_t, u_t, u_{t+1}\rangle \in R_t^1$ and $R_t^2 = \emptyset$.

Having constructed a tree T and a request R, we now proceed to prove that a uniform request $Q_d \subset R$ with load $L(Q_d) = d$ does not exist. Let $1 \le d < L^* = 2^t$. Then, d can be represented as $d = 2^{t_d}(2j_d + 1)$, where $t_d < t$ since $d < 2^t$.

Lemma 3.4 *Let $1 \le d < 2^t$, and suppose a uniform request $Q_d \subset R$ with load $L(Q_d) = d$ exists. Then, $|Q_d \cup R_i^1| \;=\; |Q_d \cup R_i^2|$ for each i in $0 \le i < t$.*

Proof: Let $|Q_d \cap R_i^1| \;=\; x_i$ and $|Q_d \cap R_i^2| \;=\; y_i$ for each i in $0 \le i \le t$. Suppose $x_i \ne y_i$ for some i. If $x_i > y_i$, in order to cover edge (u_{i+1}, v_{i+1}), the number s_{i+1} of routes in $Q_d \cap (R_{i+1}^1 \cup R_{i+1}^2)$ must be $d - y_i$, which implies that $x_i + d - y_i > d$ routes in Q_d cross edge (u_{i+1}, u_{i+2}), a contradiction of the hypothesis that $L(Q_d) = d$. Similarly, if $y_i > x_i$, the number s_{i+1} of routes in $Q_d \cap (R_{i+1}^1 \cup R_{i+1}^2)$ must be $d - x_i$ to cover edge (u_{i+1}, u_{i+2}), which implies that $y_i + d - x_i > d$ routes in Q_d cross edge (u_{i+1}, v_{i+1}), a contradiction again. Therefore, $x_i = y_i$ for each i in $0 \le i < t$. ∎

To complete the proof of Theorem 3.1, let s_i for $0 \le i \le t$ denote the number of routes in $Q_d \cap (R_i^1 \cup R_i^2)$. Then, $s_0 = d$, and Lemma 3.4 implies that $|Q_d \cap R_i^1| \;=\; |Q_d \cap R_i^2| \;=\; s_i/2$ for each $0 \le i < t$; hence, $s_i = d - s_{i-1}/2$. Recall that $d = 2^{t_d}(2j_d + 1)$ where $t_d < t$. Therefore, since the sequence of integers $(s_0, s_1, \ldots, s_{t_d}, \ldots, s_t)$ satisfies a recursion similar to the one that defines the sequence $(m_0, \ldots, m_k)$, Corollary 3.1 implies that s_{t_d} is odd. Then, since $t_d < t$, $s_{t_d+1} = d - s_{t_d}/2$ cannot be an integer, contradicting the existence of Q_d. This completes the proof of the theorem.

4 Conclusions

In this paper, we investigated the effects of inter-fiber switching capability in multi-fiber wavelength routing networks without wavelength conversion through the off-line channel assignment problem in ring and tree topologies. We characterized the switching capability of a node by its switching degree.

For a ring with inter-fiber switching at one node, and no inter-fiber switching at other nodes, we proved a bound on the load of a request that can be established, as a function of the number of fibers per link d, the number of wavelengths per link W, and the switching degree c. Shortest path routing was assumed in proving the bound. The bound is useful in evaluating the impact of limited switching capability. This bound should be compared with the best known bound of $(dW+1)/2$ for a ring with no inter-fiber switching, and the bound of dW, the maximum possible on any topology with any switching or wavelength conversion capability.

For tree topologies, we addressed the problem of determining if the best known bound of L channels per link to establish a request with load L can be met with inter-fiber switching alone. We proved that this is impossible by constructing an example.

The work here can be extended in several directions. In the tree topology, a bound of $\lfloor 3L/2 \rfloor$ channels per link is achievable with no inter-fiber switching. We have shown that the bound L is not achievable. It would be interesting to determine whether a bound between L and $\lfloor 3L/2 \rfloor$ can be achieved with inter-fiber switching alone. On-line channel assignment under limited inter-fiber switching can also be considered. Finally, other effects of inter-fiber switching, e.g., blocking performance, and other topologies can be studied.

References

[1] K. C. Lee and V. O. K. Li. A wavelength-convertible optical network. *IEEE/OSA J. Lightwave Tech.*, 11(5/6):962–970, May/June 1993.

[2] G. Jeong and E. Ayanoglu. Comparison of wavelength-interchanging and wavelength-selective cross-connects in multiwavelength all-optical networks. In *Proc. INFOCOM '96*, pages 156–163, March 1996.

[3] S. Subramaniam and R. A. Barry. Wavelength assignment in fixed routing WDM networks. In *Proc. ICC*, pages 406–410, June 1997.

[4] E. Karasan and E. Ayanoglu. Effects of wavelength routing and selection algorithms on wavelength conversion gain in WDM optical networks. *IEEE/ACM Trans. Networking*, 6(2):186–196, April 1998.

[5] X. Zhang and C. Qiao. Wavelength assignment for dynamic traffic in multi-fiber WDM ring networks. In *Proc. ICCCN*, pages 479–485, Oct. 1998.

[6] L. Li and A. K. Somani. A new analytical model for multifiber WDM networks. In *Proc. GLOBECOM*, pages 1007–1011, Dec. 1999.

[7] R. Ramaswami and G. H. Sasaki. Multiwavelength optical networks with limited wavelength conversion. *IEEE/ACM Trans. Networking*, 6(6):744–754, Dec. 1998.

[8] G. Li and R. Simha. On the wavelength assignment problem in multifiber WDM star and ring networks. In *Proc. INFOCOM*, pages 908–915, March 2000.

[9] P. Raghavan and E. Upfal. Efficient routing in all-optical networks. In *Proc. ACM Symp. Theory of Computing*, pages 134–143, May 1994.

[10] J. A. Bondy and U. S. R. Murty. *Graph Theory with Applications.* American Elsevier Publishing, 1976.

LIST OF AUTHORS

Abed-Meraim, K. 193
Abou-Faycal, I. 337
Ahmed, M. 755
Akansu, A.N. 963
Al-Dhahir, N. 680
Aleshnikov, I. 1031
Alesiani, F. 297
Al-Naffouri, T.Y. 1175
Alouini, M.-S. 42, 347, 367
Altman, E. 537
Anantharam, V. 307
Ananthaswamy, G. 377
Anigstein, P. 307
Badowski, G. 528
Balakrishnan, J. 1214
Bandyopadhyay, S. 1305
Barg, A. 81
Barton, R.J. 1
Bayoumi, M.M. 1155, 1165
Berger, T. 700
Bertsekas, D.P. 872
Bi, C. 225
Biglieri, E. 297
Black, C.K. 896
Blake, I.F. 238
Blight, D. 1266
Boariu, A. 997
Borkar, V.S. 849
Bourdon, S.E. 1145
Brehler, M. 1067
Burkykh, S. 193
Burrell, A. 577
Buzzi, S. 725, 816
Cabral, H.A. 69
Caire, G. 715
Capone, J.M. 133
Chan, A.M. 668
Charpin, P. 1029
Cheng, J. 700
Cheng, P. 926
Chindapol, A. 975
Choi, H. 1314
Choi, H.-A. 1314
Chong, E.K.P. 113, 287
Chow, P.L. 412
Chowdhury, S. 1224
Chugg, K.M. 689, 959
Corcoran, J.N. 1116
Coskun, O. 689
Costello, Jr., D.J. 69, 468
Dahleh, M.A. 916
Damen, M.O. 193
Das, D. 1194
de Farias, D.P. 855
del Angel, G. 327
Dube, P. 537
Dumer, I. 71
Dutta, R. 1284
East, J. 785
Ephremides, A. 103, 836
Erez, E. 498
Fan, J.L. 955
Feder, M. 498
Fekri, F. 957
Feng, G.-L. 593
Feron, E. 449, 916
Fine, T.L. 327
Fitz, M. 985
Fitz, M.P. 965, 995, 1057
Flanagan, J.L. 1244
Francis, B.A. 143
Frazzoli, E. 916
Frey, B.J. 62
Fuhrmann, D.R. 648
Gajic, Z. 393
Ganesan, A. 191
Garg, A. 1256
Gelfand, S.B. 357
Gentry, S. 449
Georghiades, C.N. 636, 886, 967
Giannakis, G.B. 52, 1076
Goeckel, D.L. 377
Goldsmith, A. 745
Goldstein, J.S. 1224
Goodwine, B. 906
Gosavi, A. 865
Grossglauser, M. 735
Guess, T. 795, 806
Gulliver, T.A. 258
Gupta, R. 785
Guruswami, V. 603
Gutierrez, F.J. 1294
Hagenauer, J. 1234
Hassibi, B. 1047
Hastad, J. 603
Heath, Jr., R.W. 212, 1086
Henderson, S.G. 1125
Hero, A. 785
Hochwald, B.M. 1047
Høst-Madsen, A. 32
Hottinen, A. 997
Hou, J. 266
Howitt, I. 847
Hui, J.Y. 222, 225
Huisinga, W. 1106
Ishii, H. 143
Ji, T. 961
Johansson, M. 508
Johnson, Jr., C.R. 1214
Jost, R. 133
Kalyanasundaram, S. 113
Khammash, M.H. 163
Khasminskii, R. 412
Kieffer, J. 480
Kim, H.S. 557
Knap, S. 1155, 1165

Koetter, R. 625
Koksal, A.E. 872
Kontoyiannis, I. 461
Kumar, P.V. 845, 1031
Kurzhanski, A.B. 403
Kwong, R.H. 1135
Lafortune, S. 785
Lam, W.H. 387
Laneman, J.N. 775
Lanterman, A.D. 648
Lanzetta, G. 69
Lau, W.C. 347
LaValle, S.M. 926
Lawford, M. 1145
Le Martret, C. 52
Lee, E.B. 439
Li, H. 1000
Li, J. 1000
Li, Y. 836
Liang, P. 785
Liberty, S.R. 422
Lim, J.-H. 357
Liu, Y. 995
Liu, Z. 1076
Lops, M. 816
Lott, C.G. 123
Lu, H.F.F. 845
Lynch, K.M. 896
Madhow, U. 337
Mandjes, M. 567
Mannor, S. 470
McCloud, M.L. 1067
McLaughlin, S.W. 957
Médard, M. 337, 583
Mersereau, R.M. 957
Milstein, L.B. 266
Mina, M. 1274
Mitra, U. 22
Mitter, S. 83, 93, 182, 451
Morishima, S. 1247
Mou, L. 422
Müller, R.R. 1096
Murphy, R.J. 646
Naguib, A.F. 680
Nassar, C.R. 203
Neely, M.J. 547
Nguyen, G.D. 103
Ni, J. 1
O'Sullivan, J.A. 646
Oklander, B. 488
Olshevsky, V. 613
Onggosanusi, E.N. 228
Paganini, F. 153
Papadopoulos, H. 1184
Papantoni-Kazakos, P. 577
Pauciullo, A. 816
Paulraj, A. 212, 1086
Pavlović, V. 1256
Pham, K.D. 422
Pi, Z. 22
Pless, V. 1010
Poor, H.V. 725
Pottie, G. 755
Qi, X. 163
Qian, L. 393
Raghothaman, B. 997
Rankin, D. 258
Rehg, J.M. 1256
Ricci, G. 816
Richa, A.W. 1305
Richie, J. 847
Rieken, D.W. 648
Ritcey, J.A. 975
Rohrs, C.E. 547
Rosenthal, J. 248
Rouskas, G.N. 1284
Saengudomlert, P. 583
Sahai, A. 83
Sain, M.K. 422
Salapaka, M.V. 163
Sandhu, S. 1086
Saniee, I. 567
Sasaki, G.H. 1266
Sason, I. 276
Sayed, A. 1175
Sayeed, A.M. 191, 228, 705
Schafer, R.W. 957
Schniter, P. 1204
Schütte, C. 1106
Seidl, N. 847
Sen, A. 133, 1305
Shabunov, K. 71
Shamai, S. 715
Shamir, G.I. 468
Shattil, S. 203
Shen, B.H. 1305
Shen, Z. 926
Shepard, T. 1155, 1165
Shi, Y.-Q. 945
Shokrollahi, A. 613
Shroff, N.B. 113, 557
Shum, K. 1031
Siegel, P.H. 266
Simon, M.K. 42, 347
Siwamogsatham, S. 1057
Snyder, D.L. 646
Somani, A.K. 1274
Spasojević, P. 636, 886
Stamoulis, A. 1076
Stark, W. 785
Stark, W.E. 961
Stichtenoth, H. 1031
Stolyar, A.L. 567
Su, C.-F. 1266
Subramaniam, S. 1314
Sudan, M. 603
Sun, H. 225
Sun, J. 965
Sun, Y. 826
Surijan, E. 488

Sutskover, I. 488
Takeshita, O.Y. 965, 995
Taricco, G. 297
Tatikonda, S. 93, 182, 451
Telatar, E. 276
Teneketzis, D. 123, 785
Thiennviboon, P. 959
Tilbury, D.M. 172
Tirkkonen, O. 997
Tong, L. 765
Toumpis, S. 745
Tse, D. 735
Tu, Y.-S. 755
Tüchler, M. 1234
Tulino, A.M. 11
Tweedie, R.L. 1116, 1125
Ulukus, S. 317
Urbanke, R. 276
Uysal, M. 967
Van Roy, B. 855
Van Veen, B.D. 228
Varaiya, P. 403
Varanasi, M.K. 795, 1067, 1194
Vardy, A. 625
Varvarigos, E.A. 1294
Vassiliadis, S. 1294
Venkatesh, S. 449
Verdú, S. 11
Viswanathan, H. 658
Viterbo, E. 297
Vontobel, P.O. 248
Walker, J.L. 1019
Wang, H. 785
Wang, X. 32
Wasserman, K.M. 518
Wei, X. 963
Wei, Y. 906
Weissman, T. 470
Wieselthier, J.E. 103
Wonham, W.M. 1135, 1145
Wood, J.A. 1038
Wornell, G.W. 668, 775
Worthen, A. 785
Xiao, L. 508
Yang, E.-H. 480
Yang, H. 367
Yates, R.D. 317
Yaz, E.E. 431
Yaz, Y.I. 431
Yin, G. 528
Yook, J.K. 172
You, K.H. 439
Zad, S.H. 1135
Žefran, M. 936
Zekavat, S.A. 203
Zémor, G. 81
Zhang, D. 518
Zhang, J. 287
Zhang, Q. 528
Zhang, X. 945, 985
Zhao, Q. 765
Zheng, Z.H. 387
Zhou, S. 52
Ziv, J. 488
Zoltowski, M.D. 1224
Zuckerman, D. 603

3 0112 050707634